A Separation theorem for Chain Event Graphs

Peter Thwaites[1] and Jim Q. Smith[2]
[1]University of Leeds   [2]University of Warwick

Abstract: Bayesian Networks (BNs) are popular graphical models for the representation of statistical problems embodying dependence relationships between a number of variables. Much of this popularity is due to the d-separation theorem of Pearl and Lauritzen, which allows an analyst to identify the conditional independence statements that a model of the problem embodies using only the topology of the graph. However for many problems the complete model dependence structure cannot be depicted by a BN. The Chain Event Graph (CEG) was introduced for these types of problem. In this paper we introduce a separation theorem for CEGs, analogous to the d-separation theorem for BNs, which likewise allows an analyst to identify the conditional independence structure of their model from the topology of the graph.

1 Introduction

If the DAG (directed acyclic graph) of a Bayesian Network (BN) has a vertex set $\{X_1, X_2, \ldots, X_n\}$, then there are n conditional independence assertions which can simply be read off the graph. These are the properties that state that a vertex-variable is independent of its non-descendants given its parents (the directed local Markov property [20]). Answering most conditional independence queries however, is not so straightforward. The d-separation theorem for BNs was first proved by Verma and Pearl [43], and an alternative version considered in [22, 20, 9]. The theorem addresses whether the conditional independence query $A \amalg B \mid C$? can be answered from the topology of the DAG of a BN, where A, B, C are disjoint subsets of the set of vertex-variables of the DAG. Separation theorems have been proved for more general classes of graphical model including chain graphs [6], alternative chain graphs [2], and ancestral graphs [30].

However, for many problems the available quantitative dependence information cannot all be embodied in the DAG of a BN. The Chain Event Graph

(CEG) was introduced in 2006 [39, 38] for the representation & analysis of precisely these sorts of problems. There have been a dozen papers on CEGs published since then, principally concerned with their use for problem representation (eg. [34]), probability propagation [41], learning and model selection [14, 15, 3, 33, 4, 10], and causal analysis [42, 37]. The motivation for the development of this class is that CEGs are probably the most natural graphical models for discrete processes when elicitation involves questions about how situations might unfold. Although the topology of these graphs is more complicated than that of the BN, they are more expressive, as they allow us to represent all structural quantitative information within the graph itself. Context-specific symmetries which are not intrinsic to the structure of the BN [7, 23, 29, 31] are fully expressed in the topology of the CEG, which also recognises logical or structural zeros in probability tables, and the numbers of levels taken by problem-variables. This last has been found to be essential to understanding the geometry of BN models with hidden variables [1, 26]. In this paper we return to the mathematics underpinning CEG models, and provide a separation theorem for these graphs.

The CEG is a tree-based graphical structure with a passing resemblance to graphs such as Bozga & Malers' probabilistic decision graph [8] (made popular by Jaeger et al in [17]). It differs from these in that edges in a CEG label events that might happen to an individual in a population given a particular partial history, and the coalescing of vertices & colouring of edges together encode conditional independence/Markov structure. The colouring of CEGs and their acyclicity also distinguish them from Markov state space diagrams. Finite CEGs as discussed in this paper also have finite event spaces whose atoms correspond to the distinct possible histories or developments that individuals in a population might have. The tree-structure imparts to these atoms an additional longitudinal element consisting of the stages of an individual's development. We note in passing that colour has recently been found to provide a valuable embellishment to other graphical models (see for example [16]).

Even more so than is the case with BNs, there are a number of conditional independence properties which can simply be *read* off the CEG [34], and given the tree-based nature of the CEG these properties are naturally context-specific. That is to say they are properties of the form $A \amalg B \mid \Lambda$ for some event Λ. An example would be that a particular lifestyle-related medical condition is independent of gender given that the subject is a smoker. An analogous statement for a discrete BN would be of the form

$$p(A \mid B, C = \boldsymbol{c}) = p(A \mid C = \boldsymbol{c})$$

for some subsets of variables A, B, C, some **specific** vector value $\boldsymbol{c}$ of C and all vector values of A and B. The class of conditioning events we can tackle with a CEG is however much richer than that generally considered when using BN-based analysis.

In Section 2 we use a toy example to introduce CEGs. A naive criticism of tree-based graphical structures is that they will be too complex for larger

2

problems. We note that the *picture* is simply for the investigator's (or a client's) benefit: as with any large system, analysts need to consider both local and global aspects – the full CEG for a large problem may exist only as a set of computer constraints; local aspects of the problem can be drawn out as a simple graph. Our example here is small so that we can use it effectively to illustrate key ideas.

We then formally define a CEG, and explain how coalescence & colouring encode conditional independence structure. Events & random variables defined on CEGs are introduced through our example, as are sub-CEGs conditioned on an event of interest.

In Section 3 we introduce elementary variables associated with the vertices of the CEG, and use these to construct a separation theorem. Comparison with the d-separation theorem for BNs is made through corollaries and our running example. Section 4 develops some of the ideas from earlier sections.

2 Chain Event Graphs

Definitions of Chain Event Graphs (of varying degrees of complexity) have appeared in many of the previous papers on these graphs. We offer a detailed formal definition here so that the theorems in later sections have a firm mathematical foundation.

2.1 Event Trees

We introduce CEGs in this section through the use of a toy example, simple enough to illustrate the key ideas.

The CEG is a function of an *event tree* [32], and was created to overcome some of the shortcomings of these graphs. So we start by considering an event tree elicited from some expert.

Example 1. *A researcher is investigating a population of people whose parents suffered from an inherited medical condition* C. *She has information on the gender of each individual; if and when they displayed a symptom* S *(never, before puberty, after puberty); and whether or not they developed the condition* C. *Her current research is retrospective so she also has these individual's ages at death. She suspects that the condition can lead to early death, so she produces an indicator that for each individual records whether or not they died before the age of 50. She knows that an individual who does not display symptom* S *will not develop the condition.*

An event tree for this information is given in Figure 1.

The tree is a natural description of the problem. The labels on the edges of each root-to-leaf path (eg. *male, display* S *before puberty, develop* C*, die before 50*) in Figure 1 follow a temporal order, and the absence of *condition* edges in the 9th, 10th, 19th & 20th such paths reflects the expert's knowledge that individuals who do not display S do not develop C.

Figure 1: Event tree for Example 1

An event tree $\mathcal{T}$ is a connected directed graph with no cycles. It has one root vertex (v_0) with no parents, whilst all other vertices have exactly one parent. A leaf-vertex is a vertex with no children. We denote the vertex set of $\mathcal{T}$ by $V(\mathcal{T})$ and the edge set by $E(\mathcal{T})$. A directed root-to-leaf path in $\mathcal{T}$ is called a *route*.

Although an event tree could be used to represent an observer's beliefs about the possible developments of some individual, we make the assumption that the tree relates to a population. Hence the routes of the tree describe precisely the possible developments or histories that an individual in the population can experience. This description takes the form of a sequence of edge-labels, each describing what can happen next at a vertex. So in Figure 1 for example, an individual who is male reaches vertex v_1 where the possible immediate developments are that he displays S before puberty, after puberty or not at all.

We specify that the edges leaving any vertex in the tree must have distinct labels; that each individual can only pass along one edge leaving any vertex, and the choice of which edge is determined only by the variable controlling the next stage of development (eg. *symptom*) and not by any possible further developments downstream of these edges (ie. towards the leaves).

We also require that each route corresponds to a real possible development or history of an individual in the population. So each such path has a non-zero probability that some individual might take this path. Also, the number

of routes corresponds exactly to the number of distinct possible histories or developments (defined by the edge-labels) that some individual could experience.

Once we have a set of routes, and some ordering on these paths, then the edge-labels **define** the tree structure. In our example the first variable in our order is *gender*, so v_0 has two emanating edges, labelled *male* & *female*. The second variable is *symptom*, so v_1 & v_2 both have three emanating edges, labelled *displays* S *before puberty, displays* S *after puberty* & *never displays* S. We know that individuals who never display S will not develop condition C, so the edges emanating from v_5 & v_8 label the possible values of the life-expectancy indicator, whereas those emanating from v_3, v_4, v_6 & v_7 label the possible values of *condition*.

In this paper we use the notation λ to denote a route, and the set of routes of $\mathcal{T}$ is labelled $\Lambda(\mathcal{T})$. When the tree is applied to a population, each route λ corresponds to a possible history or development of an individual in the population, and hence to an atom in an event space defined by the tree. The sigma field of events associated with $\mathcal{T}$ is then the set of all possible unions of atoms λ in $\Lambda(\mathcal{T})$. Note that the tree encodes an additional longitudinal development or history for the individual, not encoded by the sigma field alone [32]. Events in the sigma field of the tree are denoted Λ.

So for instance, in Example 1 the event Λ corresponding to *displayed* S *before puberty and died before the age of 50* is simply the union of the 1st, 3rd, 11th & 13th routes in the tree in Figure 1.

Example 1 continued. *Our researcher has done sufficient analysis of the data to tell us that:*

- *life expectancy of individuals in this population is independent of gender given that S is not displayed,*
- *males who display S at any point and females who display S before puberty have the same joint probability distribution over the variables* condition *and* life expectancy.

Moreover

- *if and when an individual displays* S *is independent of gender,*

and she believes that

- *males and females who display* S *at any point have the same probability of developing the condition.*

It is the fact that traditional trees cannot readily depict this sort of information which has led to tree-based analysis not receiving the attention it deserves. It is actually relatively easy to portray these types of conditional independence or Markov properties on a tree – all we need to do is add colour to the edges, as in Figure 2 (where edges with the same colouring carry the same probability).

Despite the colouring this is still a rather cumbersome representation. To make it more compact we use the idea of coalesced trees, used in decision analysis since [27]. In a coalesced event tree vertices from which the sets of possible complete future developments have the same probability distribution are coalesced.
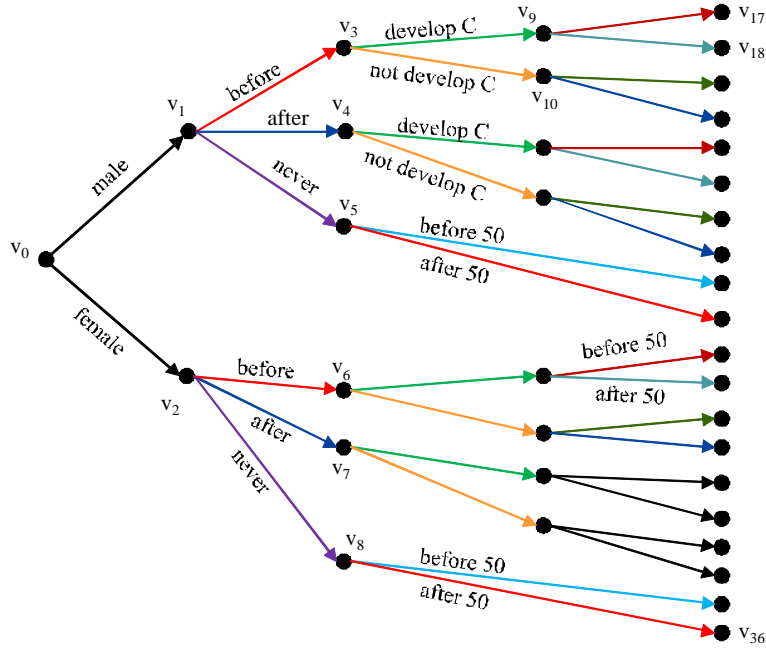
Figure 2: Coloured event tree for Example 1

So in the tree in Figure 2 we can coalesce the vertices v_3, v_4 & v_6 and also the vertices v_5 & v_8 (the vertices v_9, v_{11} & v_{13} and v_{10}, v_{12} & v_{14} are also coalesced, but this coalescence is in a sense absorbed into that of v_3, v_4 & v_6).

The combination of colouring and coalescence gives us a more compact graph that allows us to portray all conditional independence properties of the type described in Example 1 above.

So in Figure 3, the first two of the four statements provided by our researcher are depicted by the coalescence, but the latter two require the colouring of the edges leaving v_1 & v_2 and v_{3+4+6} & v_7. The colouring of the edges emanating from $v_{9+11+13}$ & from $v_{10+12+14}$ is suppressed as it no longer yields any extra information.

2.2   Probabilities on Trees

In section 2.1 we talked in general terms about probabilities on trees. In this section we formalise these ideas.

In Figure 1 the probability of the atom {*male, displayed S before puberty, de-*
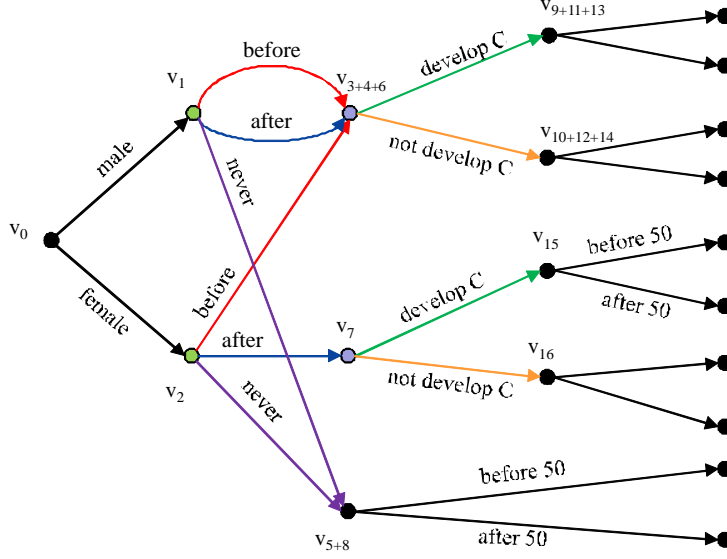
Figure 3: Coalesced event tree for Example 1

veloped C, died before 50} is clearly:

p(male) $\times$ p(displayed S before puberty | male)

$\times p$(developed C | male, displayed S before puberty)

$\times p$(died before 50 | male, displayed S before puberty, developed C)

which can be written as

$$\pi_e(v_1 \mid v_0) \; \pi_e(v_3 \mid v_1) \; \pi_e(v_9 \mid v_3) \; \pi_e(v_{17} \mid v_9)$$

where $\pi_e(v_{17} \mid v_9)$ is the probability of an individual having reached the vertex v_9 in Figure 1 (ie. they are male, displayed S before puberty & developed C) then taking the edge $e(v_9, v_{17})$ to reach the vertex v_{17} (ie. they die before 50) etc.

We can assign a probability to each atom of the event space as:

$$p(\lambda) = \prod_{e(v,v')\in\lambda} \pi_e(v' \mid v)$$

where $e(v, v')$ means the edge from vertex v to vertex v', $e(v, v') \in \lambda$ means that $e(v, v')$ lies on the route λ, and $\pi_e(v' \mid v)$ is the conditional probability of traversing the edge $e(v, v')$ given that have reached the vertex v.

We call the probabilities $\{\pi_e(v' \mid v)\}$ the *primitive* probabilities of the tree $\mathcal{T}$.

7

The set $\{p(\lambda)\}$ defines a probability measure over the sigma field of events formed by the atoms $\lambda \in \Lambda(\mathcal{T})$. Strictly speaking these probabilities are the fundamental probabilities of the system as they are the probabilities of the atoms. Each primitive probability has then a unique value determined by these $\{p(\lambda)\}$. The conditional probability of an edge $e(v, v')$ is given by:

$$\pi_e(v' \mid v) = \frac{\sum_{\lambda : e(v,v') \in \lambda} p(\lambda)}{\sum_{\lambda : v \in \lambda} p(\lambda)}$$

where the numerator is the sum of the probabilities of all routes utilising the edge $e(v, v')$, and the denominator is the sum of the probabilities of all routes passing through the *start-vertex* v of the edge $e(v, v')$. So for example in Figure 1 we have:

$$\pi_e(v_{17} \mid v_9) = \frac{p(\lambda_1)}{p(\lambda_1) + p(\lambda_2)}$$

where λ_1 is the atom corresponding to the route $v_0 \longrightarrow v_{17}$ ($\lambda(v_0, v_{17})$) and λ_2 is the atom corresponding to the route $v_0 \longrightarrow v_{18}$ ($\lambda(v_0, v_{18})$).

In practice however, our elicitation of the tree is likely to yield primitive probabilities of the sort described above, rather than probabilities of atoms.

2.3 Positions and Stages

To allow our event tree to encode the full conditional independence structure of the model we introduce two partitions of the tree's vertices.

Let $V^0(\mathcal{T})$ ($\subset V(\mathcal{T})$) be the set of non-leaf vertices of $\mathcal{T}$ (called *situations* in [42]). Also let $v \prec v'$ denote that the vertex v precedes the vertex v' on some route.

Then for any non-leaf vertex $v_a \in V^0(\mathcal{T})$ and leaf-vertex $v_a'' \in V(\mathcal{T}) \setminus V^0(\mathcal{T})$ such that $v_a \prec v_a''$, there is a unique subpath $\mu(v_a, v_a'')$ comprising of the edges of the route $\lambda(v_0, v_a'')$ which lie between the vertices v_a and v_a''.

Let:

$$\pi_\mu(v_a'' \mid v_a) = \prod_{e(v,v') \in \mu(v_a, v_a'')} \pi_e(v' \mid v)$$

Now each vertex $v \in V^0(\mathcal{T})$ labels a random variable $J(v)$ whose state space $\mathbb{J}(v)$ can be identified with the set of $v \longrightarrow leaf$ subpaths $\{\mu(v, v'')\}$.

Definition 1. Positions. *For an Event Tree $\mathcal{T}(V(\mathcal{T}), E(\mathcal{T}))$, the set $V^0(\mathcal{T})$ is partitioned into equivalence classes, called* positions *as follows:*
Vertices $v_a, v_b \in V^0(\mathcal{T})$ are members of the same equivalence class (position) *if there is a bijection ϕ between $\mathbb{J}(v_a)$ and $\mathbb{J}(v_b)$ such that if $\phi : \mu(v_a, v_a'') \mapsto \mu(v_b, v_b'')$, then*
(a) the ordered sequence of edge-labels is identical for $\mu(v_a, v_a'')$ and for $\mu(v_b, v_b'')$,
(b) $\pi_\mu(v_a'' \mid v_a) = \pi_\mu(v_b'' \mid v_b)$.

Now, from section 2.1, tree structure is defined by the edge-labels, so (a) above means that the subtrees rooted in v_a and v_b have identical topological structure.

Similarly, from section 2.2, our edge probabilities are uniquely defined by the route probabilities. We can see that the edge probabilities in the subtrees rooted in v_a and v_b must be uniquely defined by the sets of probabilities $\{\pi_\mu(v_a'' \mid v_a)\}$ and $\{\pi_\mu(v_b'' \mid v_b)\}$. So (b) above means that the corresponding edge probabilities in these two subtrees are equal.

So two vertices in a tree are in the same *position* if the sets of possible complete future developments from these vertices have the same probability distribution. We denote the set of positions of $\mathcal{T}$ by $P(\mathcal{T})$

We noted earlier that knowing this partition of vertices is insufficient for us to fully describe the conditional independence structure of the tree, so we introduce a second partition.

Each vertex $v \in V^0(\mathcal{T})$ also labels a random variable $K(v)$ whose state space $\mathbb{K}(v)$ can be identified with the set of directed edges $e(v, v')$ emanating from v.

Definition 2. Stages. *For an Event Tree $\mathcal{T}(V(\mathcal{T}), E(\mathcal{T}))$, the set $V^0(\mathcal{T})$ is partitioned into equivalence classes, called* stages *as follows:*
Vertices $v_a, v_b \in V^0(\mathcal{T})$ are members of the same equivalence class (stage) if there is a bijection ψ between $\mathbb{K}(v_a)$ and $\mathbb{K}(v_b)$ such that if $\psi : e(v_a, v_a') \mapsto e(v_b, v_b')$, then $\pi_e(v_a' \mid v_a) = \pi_e(v_b' \mid v_b)$.

So two vertices in a tree are in the same *stage* if their sets of emanating edges have the same probability distribution.

Note that the set of stages is coarser than the set of positions, and that vertices in the same position are necessarily in the same stage.

We also add colouring to trees to illustrate the stage structure. So vertices in the same stage are given the same colour, and edges emanating from vertices in the same stage are coloured according to their probabilities / labels. This induces a partition on $E(\mathcal{T})$.

2.4   Chain Event Graphs

The *Chain Event Graph $\mathcal{C}$* is a directed acyclic graph (DAG), which is connected, having a unique *root vertex* (with no incoming edges) and a unique *sink vertex* (with no outgoing edges). Unlike the BN more than one edge can exist between two vertices of a CEG. The CEG also generally has its vertices and edges coloured, although most of this paper will deal with uncoloured versions called *Simple* CEGs. The root and sink vertices of a CEG are labelled w_0 and w_∞.

Definition 3. Chain Event Graph. *The CEG $\mathcal{C}(\mathcal{T})$ (a function of the tree $\mathcal{T}(V(\mathcal{T}), E(\mathcal{T}))$) is the graph with vertex set $V(\mathcal{C})$ and edge set $E(\mathcal{C})$ defined by:*

1. *$V(\mathcal{C}) \equiv P(\mathcal{T}) \cup \{w_\infty\}$;*

2. (a) *For $w, w' \in V(\mathcal{C}) \backslash \{w_\infty\}$ there is a directed edge $e(w, w') \in E(\mathcal{C})$ iff there are vertices $v, v' \in V^0(\mathcal{T})$ such that the vertex v is in the position w ($\in P(\mathcal{T})$), v' is in the position w' ($\in P(\mathcal{T})$), and there is an edge $e(v, v') \in E(\mathcal{T})$;*
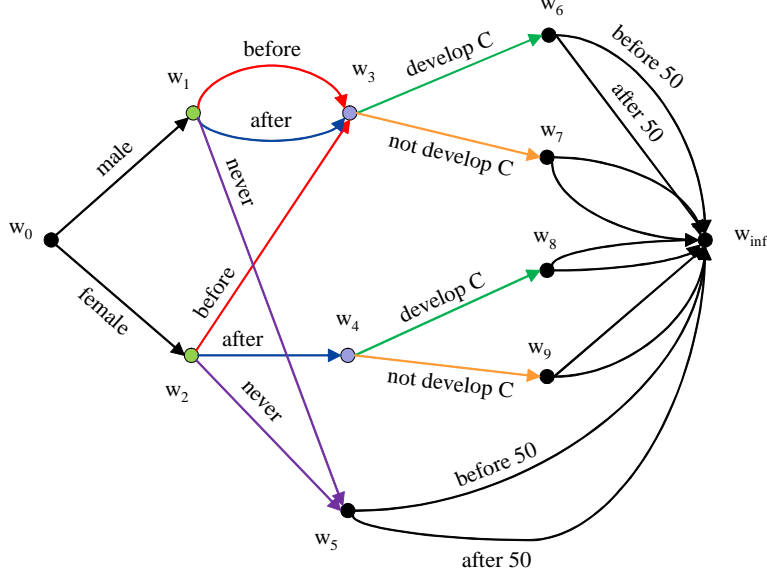
Figure 4: CEG for Example 1

(b) For $w \in V(\mathcal{C})\backslash\{w_\infty\}$ there is a directed edge $e(w, w_\infty) \in E(\mathcal{C})$ iff there is a vertex $v \in V^0(\mathcal{T})$ such that v is in the position w ($\in P(\mathcal{T})$), and there is an edge $e(v, v') \in E(\mathcal{T})$ for some leaf-vertex $v' \in V(\mathcal{T})\backslash V^0(\mathcal{T})$.

Note that the vertex set of $\mathcal{C}(\mathcal{T})$ consists of the *positions* of $\mathcal{T}$ and the sink-vertex w_∞. Positions in $\mathcal{C}(\mathcal{T})$ are said to be in the same *stage* if the component vertices (in $\mathcal{T}$) of these positions are in the same stage. Colouring in $\mathcal{C}(\mathcal{T})$ is inherited from $\mathcal{T}$. The constraints associated with the positions & stages of a CEG hold for the entire population to which the CEG has been applied,

Example 1 continued. *To convert the coalesced tree from Figure 3 to a CEG is straightforward. We simply combine the leaf-vertices into a sink-vertex w_∞ as in Figure 4.*

The positions here are w_0 through w_9. We note that w_1 & w_2 are in the same stage (as can be seen from the colouring), w_3 & w_4 are in the same stage, and each of w_5 through w_9 is in a stage by itself.

The position w_3 encodes the conditional independence / Markov property that males who display S *at any point and females who display* S *before puberty have the same joint probability distribution over the variables* condition *and* life expectancy. *The position w_5 encodes the property that life expectancy of individuals in this population is independent of gender given that* S *is not displayed. The stage $\{w_1, w_2\}$ encodes the property that if and when an individual dis-*

10

plays S *is independent of gender. The stage* $\{w_3, w_4\}$ *encodes the property that* condition *is independent of gender given* S *displayed.*

Without ambiguity we simplify our notation $\mathcal{C}(\mathcal{T})$ to $\mathcal{C}$.

Analogously with the tree, a directed $w_0 \rightarrow w_\infty$ path λ in $\mathcal{C}$ is called a *route*. The set of routes of $\mathcal{C}$ is labelled $\Lambda(\mathcal{C})$. We write $w \prec w'$ when the position w precedes the position w' on a route.

When the CEG is applied to a population, each route λ corresponds to a possible history or development of an individual in the population, and hence to an atom in the event space defined by the CEG. The sigma field of events associated with $\mathcal{C}$ is then the set of all possible unions of atoms λ in $\Lambda(\mathcal{C})$. Like the tree, the CEG encodes an additional longitudinal development or history for the individual, not encoded by the sigma field alone. Events in the sigma field of the CEG are denoted Λ.

Note that the number of routes in the CEG equals the number in the tree and corresponds exactly to the number of possible distinct histories or developments that some individual in the population could experience. And since no route has a zero probability, all edges in the CEG have non-zero conditional probabilities associated with them.

Because the CEG's atoms have this implicit longitudinal development associated with them, certain events in the sigma field are particularly important. Let $\Lambda(w)$ denote the event that an individual unit takes a route that passes through the position $w \in V(\mathcal{C})$. $\Lambda(w, w')$ is then the union of all routes passing through the positions w and w', $\Lambda(e(w, w'))$ is the union of all routes passing through the edge $e(w, w')$, and $\Lambda(\mu(w, w'))$ is the union of all routes utilising the subpath $\mu(w, w')$.

2.5 Probabilities on CEGs

As with trees, underlying the CEG there is a probability space which is specified by assigning probabilities to the atoms. For each position $w \in V(\mathcal{C})\backslash\{w_\infty\}$ and edge $e(w, w')$ emanating from w, we denote by $\pi_e(w' \mid w)$ the probability of traversing the edge $e(w, w')$ conditional on having reached the position w. We call the probabilities $\{\pi_e(w' \mid w) : e(w, w') \in E(\mathcal{C}), w \in V(\mathcal{C})\backslash\{w_\infty\}\}$ the *primitive probabilities* of $\mathcal{C}$.

Then, analogously with trees, for each atom λ:

$$p(\lambda) = \prod_{e(w,w')\in\lambda} \pi_e(w' \mid w)$$

as both the atoms and the primitive probabilities are identical to the corresponding atoms and primitive probabilities in the tree.

The set $\{p(\lambda)\}$ defines a probability measure over the sigma field of events formed by the atoms $\lambda \in \Lambda(\mathcal{C})$.

This assignment of probabilities implicitly demands a Markov property over the flow of the units through the graph. Thus, in the context of our medical example, the probablility of an individual with attributes (*male, displayed*

symptom S before puberty), (*male, displayed symptom S after puberty*) or (*female, displayed symptom S before puberty*) developing the condition depends only on the fact that the subpaths corresponding to these pairs of attributes terminate at the position w_3, and not on the particular subpath leading to w_3. The probability this individual develops the condition is then $\pi_e(w_6 \mid w_3) \equiv p(\Lambda(e(w_3, w_6)) \mid \Lambda(w_3))$. So we only need to know the position a unit has reached in order to predict as well as is possible what the next unfolding of its development will be.

This Markov hypothesis looks strong but in fact holds for many families of statistical model. For example all event tree descriptions of a problem satisfy this property, all finite state space context specific Bayesian Networks as well as many other structures [34].

We can go further and state that the sets of possible future developments (whether or not they developed the condition **and** whether or not they died before the age of 50) for individuals taking any of these three subpaths must be the same. Moreover the conditional probability of any particular subsequent development must be the same for individuals taking any of these three subpaths.

Note also that if positions w_a and w_b are such that the sets of possible future developments from w_a and w_b are identical, and the conditional joint probability distributions over these sets are identical, then w_a and w_b are the **same** position, and must be coalesced for our graph to be a CEG.

The probability of any event Λ in the sigma field is hence of the form

$$p(\Lambda) = \sum_{\lambda \in \Lambda} p(\lambda) = \sum_{\lambda \in \Lambda} \prod_{e(w, w') \in \lambda} \pi_e(w' \mid w)$$

where $\lambda \in \Lambda$ means that λ is one of the component atoms of the event Λ. In this paper we will also use the following further notation:
$\pi_\mu(w' \mid w) \equiv p(\Lambda(\mu(w, w')) \mid \Lambda(w))$ denotes the probability of utilising the subpath $\mu(w, w')$ (conditional on passing through w),
$\pi(w' \mid w) \equiv p(\Lambda(w, w') \mid \Lambda(w)) = \sum_\mu \pi_\mu(w' \mid w)$ denotes the probability of arriving at w' conditional on passing through w.

Expressing a problem as a CEG allows domain experts to check their beliefs in a very straightforward manner:

We stated in Section 2.1 that the expert in our example believed that males & females who display S at any point have the same probability of developing C. This is depicted in the colouring of the edges emanating from w_3 & w_4 in Figure 4. Our expert can now use the techniques developed in [14, 15, 3, 33] to test the model represented by Figure 4 against alternative models with different conditional independence / Markov structure. Such a test might yield information that grouped the vertices v_3, v_4, v_6 & v_7 from Figure 1 into different positions than those in Figure 4; or that the vertices v_3, v_4 & v_6 are indeed in the same stage and position, but that the vertex v_7 is not in this stage (ie. the probability of developing C is different for females who display S after puberty), and so the edges leaving w_3 & w_4 in Figure 4 would no longer have the same colouring.

2.6 Conditioning on events

Most conditional independence queries that could realistically be of interest to an analyst can be answered purely by inspecting the topology of a CEG. And most of these queries involve conditioning on what is known as an intrinsic event.

Definition 4. Intrinsic events. *Let $\mathcal{C}_\Lambda$ be the subgraph of $\mathcal{C}$ consisting of only those positions and edges that lie on a route $\lambda \in \Lambda$, and the sink-vertex w_∞. Λ is intrinsic to C if the number of $w_0 \longrightarrow w_\infty$ paths in $\mathcal{C}_\Lambda$ equals the number of atoms in the set $\{\lambda\}_{\lambda \in \Lambda}$.*

The idea of intrinsic events is closely related to that of faithfulness in BNs [24, 36]. Note that each atom (& therefore edge) in $\mathcal{C}_\Lambda$ must by construction have a non-zero probability, but edge-probabilities in $\mathcal{C}_\Lambda$ may differ from those in $\mathcal{C}$ since some vertices in $\mathcal{C}_\Lambda$ will have fewer emanating edges than they have in $\mathcal{C}$, and the probabilities on the emanating edges of any vertex must sum to one.

All atoms of the sigma field of $\mathcal{C}$ are intrinsic, as are $\Lambda(w)$, $\Lambda(w, w')$, $\Lambda(e(w, w'))$, $\Lambda(\mu(w, w'))$ (provided these are non-empty), and as is the exhaustive set $\Lambda(w_0)$. If we include the empty set in the set of intrinsic events then we note that intrinsic sets are closed under intersection and so technically form a π-system (see for example [18]) we can associate with the CEG $\mathcal{C}$.

Not all events in the sigma field are necessarily intrinsic, because the class of intrinsic events is not closed under union. For example, for the CEG in Figure 4, the event Λ consisting of the union of the two atoms described by the routes *male, display* S *before puberty, develop* C, *die before 50* and *male, display* S *after puberty, develop* C, *die after 50* produces a subgraph $\mathcal{C}_\Lambda$ which has four distinct routes, so Λ is not intrinsic. However, our interest in intrinsic events is that we can condition on them, and we show below that conditioning on intrinsic events often destroys the stage-structure of $\mathcal{C}$. Conditioning on non-intrinsic events usually destroys position-structure. From this we argue that if we know that we wish to condition on an event such as the one described above, we would simply sacrifice the position-structure of our CEG (knowing that it would probably be lost in the conditioning anyway) and split (uncoalesce) the position w_3 to form a graph for which this event is intrinsic.

Even without such sleight of hand, the class of intrinsic events is rich enough to encompass virtually all of the conditioning events in the conditional independence statements we would like to query. In particular, if our model can be expressed as a BN (with vertex-variables $\{X_j\}$) then any set of observations expressible in the form $\{X_j \in A_j\}$ ($\{A_j\}$ subsets of the sample spaces of $\{X_j\}$) is a proper subset of the set of intrinsic events defined on the CEG of our model [41].

Example 1 continued. *Suppose for illustrative convenience that the edges labelled* male, female, displayed S before puberty, displayed S after puberty, never displayed S, developed C, did not develop C *in our CEG have the probabilities* $\frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$. *Now let us condition on the event Λ which is the union of all routes* **except** {female, displayed S after puberty, did not develop C, died

before 50} *and* {female, displayed S after puberty, did not develop C, died after 50}. *This event* Λ *is clearly intrinsic to* $\mathcal{C}$*, and has the probability* $p(\Lambda) = \frac{15}{16}$.

When we condition on Λ*, the routes* λ *which are components of* Λ *get new probabilities* $p(\lambda \mid \Lambda) = p(\lambda, \Lambda)/p(\Lambda) = p(\lambda)/p(\Lambda)$*. In this case each route has its probability multiplied by* $\frac{16}{15}$*. We leave it as a (simple) exercise to show that all edge-probabilities remain unchanged except:*

$$\pi_e(w_1 \mid w_0) \text{ becomes } 8/15$$
$$\pi_e(w_2 \mid w_0) \text{ becomes } 7/15$$
$$\pi_e(w_3 \mid w_2) \text{ becomes } 2/7$$
$$\pi_e(w_4 \mid w_2) \text{ becomes } 1/7$$
$$\pi_e(w_5 \mid w_2) \text{ becomes } 4/7$$
$$\pi_e(w_8 \mid w_4) \text{ becomes } 1$$
$$\text{the edge } \pi_e(w_9 \mid w_4) \text{ does not exist in } \mathcal{C}_\Lambda$$

So the positions w_1 *and* w_2 *are no longer in the same stage.*

So, as already noted, conditioning on an intrinsic event can destroy stage-structure. This leads us to define an uncoloured version of the CEG.

Definition 5. Simple CEG *A simple CEG (sCEG) is a CEG where there are no constraints on edge-probabilities, except that (i) all edge-probabilities must be greater than zero (a consequence of the requirement we made for trees), and (ii) the sum of emanating-edge-probabilities for any position must equal one.*

What this means in practice is that stage-structure is suppressed: there are no stages which are not positions, and so colouring is redundant. There is an analogy here with BNs to which one can always **add** edges, and sacrifice a little conditional independence structure.

We show now that the class of sCEG models is closed under conditioning on an intrinsic event:

Theorem 1. *For an event* Λ*, intrinsic to* $\mathcal{C}$*, the subgraph* $\mathcal{C}_\Lambda$ *is an sCEG. If the probability of any route* λ *in the sigma field of* $\mathcal{C}_\Lambda$ *is given by* $p_\Lambda(\lambda) = p(\lambda \mid \Lambda)$*, then the edge-probabilities in* $\mathcal{C}_\Lambda$ *are given by:*

$$\hat{\pi}_e(w' \mid w) = \frac{p(\Lambda \mid \Lambda(e(w, w')))}{p(\Lambda \mid \Lambda(w))} \; \pi_e(w' \mid w)$$

The proof of this theorem is in the appendix. We note that this result has been successfully used to develop fast propagation algorithms for CEGs [41].

Note that the probability of an atom λ in $\mathcal{C}$ conditioned on the intrinsic event Λ is the probability of that atom in the sCEG $\mathcal{C}_\Lambda$ (denoted $p_\Lambda(\lambda)$). It is then trivially the case that the probability of an **event** in $\mathcal{C}$ conditioned on the event Λ is the probability of that event in the sCEG $\mathcal{C}_\Lambda$.

2.7 Random variables on sCEGs

Random variables measurable with respect to the sigma field of $\mathcal{C}$ partition the set of atoms into events. So consider a random variable X with state space $\mathbb{X}$, and let us denote the event that X takes the value x $(\in \mathbb{X})$ by Λ_x. Then the set $\{\Lambda_x\}_{x \in \mathbb{X}}$ partitions $\Lambda(\mathcal{C})$.

For any CEG there is a set of fairly transparent random variables which includes as a subset the set of measurement variables of any BN-representation of the model, if such a representation exists. These are called *cut-variables* and are discussed in detail in Section 3.2. In Figure 4 for example, we have a variable which could be called *symptom*, which could take the values 1, 2 & 3 (in some order) for routes traversing edges labelled *before*, *after* and *never*. These are not however the only variables we can define on a CEG, and we first consider some results for general variables.

Note that when we write $X \amalg Y$ we mean that $p(X = x, Y = y) = p(X = x)\, p(Y = y)\ \forall x \in \mathbb{X}, y \in \mathbb{Y}$, and that this is true for all distributions P compatible with $\mathcal{C}$. Now for an intrinsic event Λ, we can write $X \amalg Y \mid \Lambda$ if and only if $p(X = x, Y = y \mid \Lambda) = p(X = x \mid \Lambda)\, p(Y = y \mid \Lambda)$ for all values x of X and y of Y (see for example [12]). That is $X \amalg Y \mid \Lambda \Leftrightarrow p(\Lambda_x, \Lambda_y \mid \Lambda) = p(\Lambda_x \mid \Lambda)\, p(\Lambda_y \mid \Lambda)$ for all $\Lambda_x \in \{\Lambda_x\}_{x \in \mathbb{X}}$, $\Lambda_y \in \{\Lambda_y\}_{y \in \mathbb{Y}}$.

Lemma 1. *For a CEG $\mathcal{C}$, variables X, Y measurable with respect to the sigma field of $\mathcal{C}$, and intrinsic conditioning event Λ, the statement $X \amalg Y \mid \Lambda$ is true if and only if $X \amalg Y$ is true in the sCEG $\mathcal{C}_\Lambda$.*

The proof of this lemma is in the appendix. This is a particularly useful property because it allows us to check any context-specific conditional independence property by checking a non-conditional independence property on a sub-sCEG.

To motivate the theory in the remainder of section 2 and in section 3, we need a bigger example.

Example 2. *Our researcher from Example 1 now turns her attention to an ongoing study. Subjects who display the symptom* S *(at any point) may be given a drug, and the probability of receiving this drug is not dependent on their gender or when they displayed* S*. Those that develop the condition* C *may be given treatment, and the probability of receiving this treatment is not dependent on their gender, when they displayed* S*, or whether or not they received the earlier drug. The CEG for this is given in Figure 5.*

These two properties are depicted in the CEG by the positions w_3 & w_4 being in the same stage, and by w_{10}, w_{12} & w_{14} also being in the same stage.

Figure 5 also tells us that treatment and life expectancy are independent of gender and when S *displayed, given both the event* drug given & develop C *and the event* drug given & not develop C *(the positions w_{10} & w_{11}); and that life expectancy is independent of gender and when* S *displayed given the event* drug not given, develop C & treatment given *(the position w_{18}).*
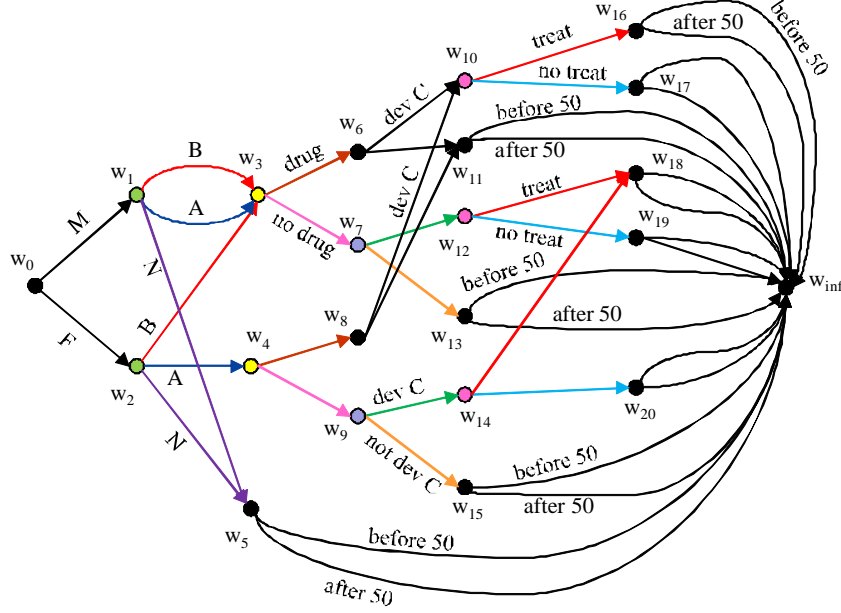
Figure 5: CEG for Example 2

Our researcher is interested in the relationships between condition & gender, and between condition & when S *displayed, for the subgroups (i) who were given the drug, and (ii) who displayed* S *but were not given the drug.*

In BN-theory, if we wish to answer the query $X \amalg Y \mid Z$?, one way we might start doing this is by drawing the ancestral graph of $\{X, Y, Z\}$ (see for example [21]). We do this because variables in the BN which are not part of this graph have no influence on the outcome of our query.

There is no direct analogy for this graph in CEG-theory, but we can consider a pseudo-ancestral graph associated with a set of events or variables. So in Example 2, all edges associated with treatment or life expectancy lie downstream (ie. towards the sink-node) of the edges associated with gender, symptom, drug and condition, so we can simply curtail our CEG so that it does not include these edges.

So in Figure 5, the positions $w_5, w_{10}, w_{11}, w_{12}, w_{13}, w_{14}$ & w_{15} are coalesced into a new sink-node w_∞ as in Figure 6. But w_7 & w_9 in Figure 5 were in the same stage. As these nodes are now only one edge upstream of w_∞, they get coalesced into a single new position (w_7 in Figure 6). Notice how much simpler the pseudo-ancestral graph is than the original CEG.

We have noted above that stage-structure is often destroyed by conditioning on an intrinsic event, but that the set of sCEGs is closed under this conditioning.
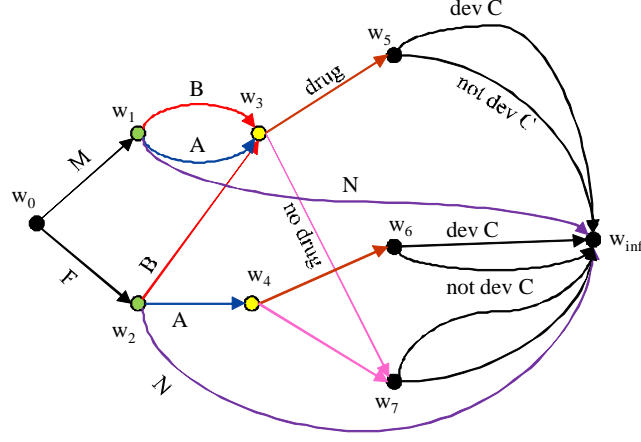
Figure 6: Pseudo-ancestral graph for Example 2

So the remainder of our analysis is conducted on an uncoloured CEG.

The graph in Figure 7 is the uncoloured pseudo-ancestral sCEG $\mathcal{C}$ associated with the queries that our researcher is interested in. This graph is analogous to the moralised ancestral graph used in BN-based analysis.

There are two natural variables which partition $\Lambda(\mathcal{C})$ – these are gender (which partitions $\Lambda(\mathcal{C})$ into events which we will call M (male) & F (female)), and symptom (which partitions $\Lambda(\mathcal{C})$ into events which we will call B (S displayed before puberty), A (S displayed after puberty) & N (S never displayed)).

The variable associated with giving the drug partitions $\Lambda(\mathcal{C})$ into three events – drug given, S displayed but drug not given, and S not displayed and hence drug not given. As the third of these events is exactly the event N above, we will for brevity describe the second event (particularly when labelling edges) simply as drug not given or no drug.

The variable associated with condition C partitions $\Lambda(\mathcal{C})$ into three events – C developed, S displayed but C not developed, and S not displayed and hence C not developed. Again, as the third of these events is exactly the event N, we will for brevity describe the second event simply as C not developed.

There is no ambiguity here as the queries our researcher is interested in correspond to conditioning on the events drug given, and S displayed but drug not given.

Her first question concerns the relationship between condition and when S displayed for the subgroup who were given the drug. This requires conditioning on the event drug given, so we draw the sub-sCEG $\mathcal{C}_\Lambda$ for this event. This is given in Figure 8.

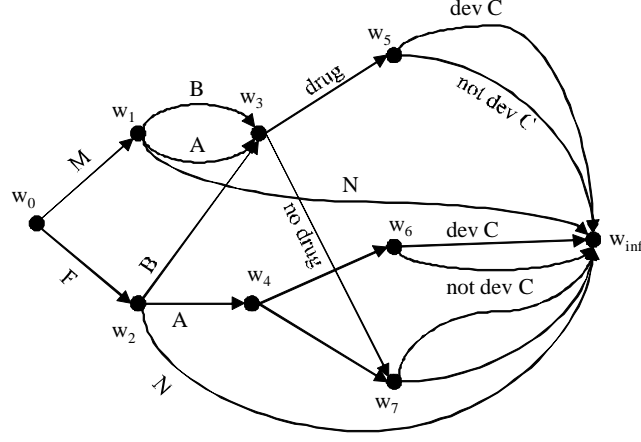The relationships she is interested in concern the probabilities $p_\Lambda(\text{C developed} \mid \text{B})$

Figure 7: Pseudo-ancestral sCEG $\mathcal{C}$ for Example 2

and $p_\Lambda(\text{C developed} \mid \text{A})$, and these are given below:

$p_\Lambda(\text{C developed} \mid \text{B})$

$= \{p_\Lambda(\text{M})\ p_\Lambda(\text{B} \mid \text{M}) \times 1 \times p_\Lambda(\text{C developed} \mid (\text{M}, \text{B})\ \text{or}\ (\text{M}, \text{A})\ \text{or}\ (\text{F}, \text{B}))$

$\qquad + p_\Lambda(\text{F})\ p_\Lambda(\text{B} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed} \mid (\text{M}, \text{B})\ \text{or}\ (\text{M}, \text{A})\ \text{or}\ (\text{F}, \text{B}))\}$

$\qquad \div \{p_\Lambda(\text{M})\ p_\Lambda(\text{B} \mid \text{M}) + p_\Lambda(\text{F})\ p_\Lambda(\text{B} \mid \text{F})\}$

$= p_\Lambda(\text{C developed} \mid (\text{M}, \text{B})\ \text{or}\ (\text{M}, \text{A})\ \text{or}\ (\text{F}, \text{B}))$

$= p(\text{C developed} \mid ((\text{M}, \text{B})\ \text{or}\ (\text{M}, \text{A})\ \text{or}\ (\text{F}, \text{B})),\ \text{drug given})$ $\qquad\qquad$ (2.1)

Note:

1. We do not need for our purposes here to evaluate the $p_\Lambda(\dots)$ probabilities, but if we wished to we could use the expression from Theorem 1.

2. The expression (2.1) is still the simplest expression even if we were to reintroduce stage-structure and let w_1 & w_2 be in the same stage.

$p_\Lambda(\text{C developed} \mid \text{A})$

$= \{p_\Lambda(\text{M})\ p_\Lambda(\text{A} \mid \text{M}) \times 1 \times p_\Lambda(\text{C developed} \mid (\text{M}, \text{B})\ \text{or}\ (\text{M}, \text{A})\ \text{or}\ (\text{F}, \text{B}))$

$\qquad + p_\Lambda(\text{F})\ p_\Lambda(\text{A} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed} \mid \text{F}, \text{A})\}$

$\qquad \div \{p_\Lambda(\text{M})\ p_\Lambda(\text{A} \mid \text{M}) + p_\Lambda(\text{F})\ p_\Lambda(\text{A} \mid \text{F})\}$

which clearly does not equal expression (2.1).

3. The denominator is of course $p_\Lambda(\text{A})$, but even if we let w_1 & w_2 be in the same stage, the above expression only simplifies to
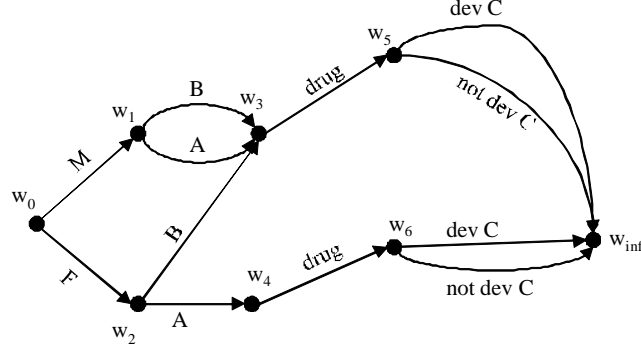
18

Figure 8: $\mathcal{C}_\Lambda$ for $\Lambda = \{\text{drug given}\}$

> $p_\Lambda(\text{M})\, p_\Lambda(\text{C developed} \mid (\text{M}, \text{B})$ or (M, A) or $(\text{F}, \text{B}))+p_\Lambda(\text{F})\, p_\Lambda(\text{C developed} \mid \text{F}, \text{A})$,
> which still does not equal expression (2.1).

Suppose we now consider the subgroup who displayed S but were not given the drug and the sub-sCEG $\mathcal{C}_\Lambda$ for the event drug not given. This CEG is given in Figure 9.

The corresponding probabilities are:

$p_\Lambda(\text{C developed} \mid \text{B})$

$$\begin{aligned}
&= \{p_\Lambda(\text{M})\, p_\Lambda(\text{B} \mid \text{M}) \times 1 \times p_\Lambda(\text{C developed}) \\
&\quad + p_\Lambda(\text{F})\, p_\Lambda(\text{B} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed})\} \\
&\quad \div \{p_\Lambda(\text{M})\, p_\Lambda(\text{B} \mid \text{M}) + p_\Lambda(\text{F})\, p_\Lambda(\text{B} \mid \text{F})\} \\
&= p_\Lambda(\text{C developed}) \quad = p(\text{C developed} \mid \text{drug not given})
\end{aligned}$$

$p_\Lambda(\text{C developed} \mid \text{A})$

$$\begin{aligned}
&= \{p_\Lambda(\text{M})\, p_\Lambda(\text{A} \mid \text{M}) \times 1 \times p_\Lambda(\text{C developed}) \\
&\quad + p_\Lambda(\text{F})\, p_\Lambda(\text{A} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed})\} \\
&\quad \div \{p_\Lambda(\text{M})\, p_\Lambda(\text{A} \mid \text{M}) + p_\Lambda(\text{F})\, p_\Lambda(\text{A} \mid \text{F})\} \\
&= p_\Lambda(\text{C developed}) \quad = p(\text{C developed} \mid \text{drug not given})
\end{aligned}$$

So we have that *whether* C *developed is independent of when* S *displayed, given that* S *was displayed but the drug was not given.*

4. We do not need to consider the case where S never displayed, as this has no intersection with Λ: Given that S *displayed but drug not given*, I **know** that S was displayed, but further knowledge of when it was displayed is irrelevant for prediction of whether or not the subject developed C,
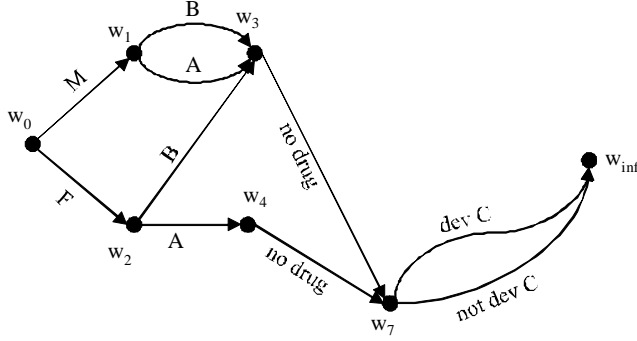
Figure 9: $\mathcal{C}_\Lambda$ for $\Lambda = \{$drug not given$\}$

5. This context-specific conditional independence property holds whether or not we reintroduce stage-structure and let w_1 & w_2 be in the same stage.

Let us now consider our researcher's other queries to do with the relationship between condition and gender for our subgroups. From Figure 8 we can see that if $\Lambda = \{$drug given$\}$ then:

$p_\Lambda(\text{C developed} \mid \text{M})$
$= p_\Lambda(\text{A or B} \mid \text{M}) \times 1 \times p_\Lambda(\text{C developed} \mid (\text{M}, \text{B}) \text{ or } (\text{M}, \text{A}) \text{ or } (\text{F}, \text{B}))$ (2.2)

and $p_\Lambda(\text{A or B} \mid \text{M}) = 1$, since A & B are the only edges leaving w_1 in $\mathcal{C}_\Lambda$.

$p_\Lambda(\text{C developed} \mid \text{F})$
$= p_\Lambda(\text{B} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed} \mid (\text{M}, \text{B}) \text{ or } (\text{M}, \text{A}) \text{ or } (\text{F}, \text{B}))$
$+ p_\Lambda(\text{A} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed} \mid \text{F}, \text{A})$

which clearly does not equal expression (2.2), and this is true even if we reintroduce stage-structure and let w_1 & w_2 be in the same stage.

From Figure 9 we can see that if $\Lambda = \{$drug not given$\}$ then:

$p_\Lambda(\text{C developed} \mid \text{M})$
$= p_\Lambda(\text{A or B} \mid \text{M}) \times 1 \times p_\Lambda(\text{C developed})$
$= p_\Lambda(\text{C developed})$ $= p(\text{C developed} \mid \text{drug not given})$
$p_\Lambda(\text{C developed} \mid \text{F})$
$= p_\Lambda(\text{B} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed})$
$+ p_\Lambda(\text{A} \mid \text{F}) \times 1 \times p_\Lambda(\text{C developed})$
$= p_\Lambda(\text{C developed})$ $= p(\text{C developed} \mid \text{drug not given})$

20

So we have that *whether* C *developed is independent of gender, given that* S *was displayed, but the drug was not given,* and this is true irrespective of whether we reintroduce stage-structure.

We notice that the topological feature which distinguishes Figure 9 from Figure 8 is that in Figure 9 there is a cut-vertex (a single vertex not w_0 or w_∞, through which all routes in the graph pass) lying between the edges associated with gender & symptom (upstream) and those associated with condition (downstream). We return to cut-vertices and to their role in independence queries in section 3.

Note also that the example above gives ample justification for working with sCEGs when considering conditional independence queries, rather than their coloured counterparts.

3   A separation theorem for simple CEGs

In section 2.7 we introduced random variables on sCEGs. In section 3.1 we develop this idea, before providing a separation theorem for sCEGs in section 3.2.

3.1   Position variables

As noted in Section 1, modified BNs of one type or another are widely used because real problems tend to contain more symmetries than can be represented by a standard BN. What is generally not addressed in papers on these types of graphs is the consequence that this extra structure has for the Markov relationships between the problem variables. With CEGs we can address this explicitly & automatically, and the first step towards doing this is to consider model variables which are more fundamental than the measurement variables customarily considered when working with BNs. So in this section we describe two types of elementary random variables, measurable with respect to the sigma field of $\mathcal{C}$, that can be identified with each position $w \in V(\mathcal{C}) \backslash \{w_\infty\}$. These are the variables $\{I(w)\}\}$ and $\{X(w)\}$ defined below.

Note that when we say that a variable X takes the value x, this is equivalent to saying that an individual from our population has a development which we equate with a route λ, and that this route λ is an element of Λ_x, the event corresponding to $X = x$.

For a position w, $I(w)$ can take the values 1 or 0 depending on whether this individual is on a route λ which does or does not pass through w. So:

$$I(w) = \left\{ \begin{array}{ll} 1 & \text{if } w \in \lambda \\ 0 & \text{if } w \notin \lambda \end{array} \right.$$

(where as above, $w \in \lambda$ means that the position w lies on the route λ).

Up until now we have labelled edges by their start and endpoints (eg. $e(w, w')$), but we can also label the edges leaving a position w by a set of

21

arbitrary labels of the form $e_x(w)$ $(x = 1, 2, \ldots)$. We define $X(w)$ by:

$$X(w) = \begin{cases} x & \text{if } e_x(w) \in \lambda \\ 0 & \text{if } w \notin \lambda \end{cases}$$

So $X(w) = x$ $(\neq 0)$ if our individual is on a route λ which passes through w and the edge $e_x(w)$.

Recall that a CEG depicts all possible histories of a unit in a population, and gives a probability distribution over these histories. However, when a single unit traverses one of the routes in the CEG, values are assigned to $I(w)$ & $X(w)$ for all positions $w \in V(\mathcal{C})$.

Notice that since $I(w)$ is clearly a function of $X(w)$, to specify a full joint distribution over the position variables, it is sufficient to specify the joint distribution of $\{X(w) : w \in V(\mathcal{C})\backslash\{w_\infty\}\}$. Note also that all atoms λ can be expressed as an intersection:

$$\lambda = \bigcap_{w \in \lambda} \{X(w) = x_\lambda\},$$

and events in the sigma algebra of $\mathcal{C}$ as the union of these atoms:

$$\Lambda = \bigcup_{\lambda \in \Lambda} \left\{ \bigcap_{w \in \lambda} \{X(w) = x_\lambda\} \right\},$$

where x_λ $(\neq 0)$ is the unique value of $X(w)$ labelling the edge in the route λ.

Up until this point we have used the words *upstream* and *downstream* rather loosely – in the context of **sets** of **edges** we have simply used these words to mean further towards w_0 and further towards w_∞; but we need to formalise the meanings here in the context of **positions**. So when we say that w_1 is upstream of w_2, or w_2 is downstream of w_1, we mean that $w_1 \prec w_2$.

For any set $A \subset V(\mathcal{C})$, let $\boldsymbol{X}_A$ denote the set of random variables $\{X(w) : w \in A\}$ and $\boldsymbol{I}_A$ the set $\{I(w) : w \in A\}$. Also, for any $w \in V(\mathcal{C})$, let $U(w)$ be the set of positions in $V(\mathcal{C})$ which lie upstream of the position w, $D(w)$ the set of positions which lie downstream of w, $U^c(w)$ the set of positions which do not lie upstream of w, and $D^c(w)$ the set of positions which do not lie downstream of w.

Lemma 2. *For any sCEG $\mathcal{C}$ and position $w \in V(\mathcal{C})\backslash\{w_\infty\}$, the variables $I(w), X(w)$ exhibit the* position independence property *that*

$$X(w) \amalg \boldsymbol{X}_{D^c(w)} \mid I(w)$$

This result (an extension of the Limited Memory Lemma of [37]) is analogous to the Directed Markov property which can be used to define BNs (see for example [28]), and which states that a BN vertex-variable is independent of its non-descendants given its parents. It provides a set of conditional independence statements that can simply be read from the graph, one for each position in $V(\mathcal{C})$. The proof of the lemma is in the appendix.

The statement that $X(w) \amalg \boldsymbol{X}_{D^c(w)} \mid (I(w) = 1)$ can be read as: *Given a unit reaches a position $w \in V(\mathcal{C})$, whatever happens immediately after w is independent of not only all developments through which that position was reached, but also of all positions that logically have not happened or could not now happen because the unit has passed through w.*

3.2 Theorem and corollaries

It is doubtful whether BNs would have enjoyed their enormous popularity if it were not so apparently easy to read conditional independence properties from them. In particular, the existence of the d-separation theorem [22, 43] has allowed all practitioners to make some attempt at model interpretation with some degree of confidence.

The presence of any context-specific conditional independence structure however severely hampers analysts using BNs in their attempts to get accurate pictures of the structure of their problems [7, 29]. In earlier sections of this paper (and in particular in section 2.7) we have been developing the theory needed for reading and representing (context-specific) conditional independence structure using CEGs. In particular, Lemma 1 allows us to consider context-specific queries by looking at the relevant sub-CEG; and Example 2 provides the rationale for looking at sCEGs. We now provide a separation theorem for sCEGs.

Using the standard terminology of non-probabilistic graph theory, we call a position $w \in V(\mathcal{C}) \backslash \{w_\infty\}$ a *cut-vertex* if the removal of w and its associated edges from $\mathcal{C}$ would result in a graph with two disconnected components. An alternative description would be *a position other than w_0 through which all routes pass*. We also remind readers at this point that when we write (for example) $X \amalg Y$ we mean that $p(X = x, Y = y) = p(X = x) \ p(Y = y)$ $\forall x \in \mathbb{X}, y \in \mathbb{Y}$, and that this is true for all distributions P compatible with $\mathcal{C}$.

Theorem 2. *In an sCEG $\mathcal{C}$ with $w_1, w_2 \in V(C) \backslash \{w_\infty\}$ and $w_2 \nprec w_1$, $X(w_1) \amalg X(w_2)$ if and only if either (i) there exists a cut-vertex w such that $w_1 \prec w \prec w_2$, or (ii) w_2 is itself a cut-vertex.*

The proof of this theorem is in the appendix. The variables $\{I(w)\}\}$ and $\{X(w)\}$ have an obvious intrinsic mathematical interest, but for more practical purposes we need to be able to make statements about the relationships between variables which are more closely analogous to the measurement variables used in BN-based analysis. So, in the same way that our primitive probabilities were used to build probabilities of subpaths and routes, we can use the $X(w)$ variables to build new *bigger* variables which have a more transparent interpretation for the analyst.

In Figure 4, let $X(w_i)$ (for $i = 5, 6, 7, 8, 9$) equal 1 if an individual has a development which takes them through the position w_i and then they die before the age of 50, and equal 2 if they have a development which takes them through the position w_i and then they die after the age of 50. Since an individual's

23

development will take them through one & only one of $\{w_5, w_6, w_7, w_8, w_9\}$, we can define a life expectancy indicator across the whole CEG by

$$\sup_{w_i : i \in \{5,6,7,8,9\}} X(w_i)$$

which takes the value 1 if an individual dies before the age of 50, or 2 if they die after the age of 50.

Analogously with the idea of a cut-vertex, a *position cut* is a set of positions the removal of which from $V(\mathcal{C})$ would result in a graph with two disconnected components. This is formalised in Definition 6.

Definition 6. Position cut. *A set of positions $W \subset V(\mathcal{C}) \backslash \{w_0, w_\infty\}$ is a position cut if $\{\Lambda(w) : w \in W\}$ forms a partition of $\Lambda(\mathcal{C})$.*

As noted above, for any position cut W, we can define a *cut-variable*; this is formalised in Definition 7.

Definition 7. Cut-variable. *For a position cut W, the random variable $X(W) \equiv \sup_{w \in W} X(w)$ is called a* cut-variable.

Note that $X(W)$ can also be defined as $X(W) \equiv \sum_{w \in W} X(w)$. The equivalence of the two forms comes from the fact that $X(w) > 0$ for one & only one position $w \in W$.

In Figure 4 we have the obvious cut-variables *gender* and *symptom*. If we assign values of 1 to edges labelled *develop* C, 2 to *not develop* C, 3 to *die before 50*, and 4 to *die after 50*, then $X(W)$ for $W = \{w_3, w_4, w_5\}$ becomes a more sophisticated cut-variable for developing the condition: $X(W)$ takes the value 1 if & only if an individual develops C, but $X(W) = 2$ tells us that an individual displayed symptom S yet did not develop C, and $X(W) = 3$ or 4 tells us that an individual did not display S and therefore did not develop C.

Theorem 2 allows us to look at the detail of the Markov structure depicted by our CEGs. The following corollaries allow us to get a broader picture.

Corollary 1. *For an sCEG $\mathcal{C}$ with position cuts W_a and W_b, the property $X(w_1) \amalg X(w_2)$ holding for any $w_1 \in W_a$, $w_2 \in W_b$ implies that $X(W_a) \amalg X(W_b)$.*

So, as one might expect, the presence of a cut-vertex in an sCEG renders cut-variables upstream of this vertex independent of cut-variables downstream of the vertex. The proof of the corollary is in the appendix.

As already noted, CEGs have been designed for the representation and analysis of asymmetric problems; and for symmetric problems a graph such as a BN is more appropriate. But it is clear that where a problem can also be adequately represented as a BN (without too much context-specific structure), the set of cut-variables of a CEG-representation must contain the set of variables associated with the vertices of the BN, as these are simply the measurement variables of the problem. Hence, if an sCEG $\mathcal{C}$ represents a model which admits a product space structure, M, N are measurement variables of the model associated

with position cuts W_M, W_N, then the property $M \amalg N$ holds providing that $X(w_m) \amalg X(w_n)$ for any $w_m \in W_M, w_n \in W_N$. This result follows immediately from Corollary 1.

Of more interest to analysts of asymmetric problems is the result given in Corollary 2, which ties together the ideas presented in Corollary 1 and Lemma 1.

Corollary 2. *Let $\mathcal{C}$ be a CEG with position cuts W_a, W_b, and Λ an event intrinsic to $\mathcal{C}$. If, in the sCEG $\mathcal{C}_\Lambda$, there exists a cut-vertex w such that $W_a \prec w \prec W_b$, then $X(W_a) \amalg X(W_b) \mid \Lambda$.*

The proof of this corollary is in the appendix. We can immediately deduce that if a CEG $\mathcal{C}$ represents a model which admits a product space structure, M, N are measurement variables of the model associated with position cuts W_M, W_N, and Λ is an event intrinsic to $\mathcal{C}$, then if in the sCEG $\mathcal{C}_\Lambda$ there exists a cut-vertex w such that $W_M \prec w \prec W_N$, the property $M \amalg N \mid \Lambda$ must hold.

Recall from Section 2.7 that for a measurement variable X with state space $\mathbb{X}$, the event that X takes the value x ($\in \mathbb{X}$) is denoted by Λ_x, and the set $\{\Lambda_x\}_{x \in \mathbb{X}}$ partitions $\Lambda(\mathcal{C})$. So the query $M \amalg N \mid X$? can be answered by checking the queries $M \amalg N \mid \Lambda_x$? for each $x \in \mathbb{X}$. If our problem elicitation indicates that there are no context-specific variations in independence properties connected with conditioning on the variable X, we can answer the query $M \amalg N \mid X$? by looking at a single graph $\mathcal{C}_{\Lambda_x}$ for some convenient value $X = x$.

Moreover, although this argument has been constructed under the assumptions that $\mathcal{C}$ admits a product space structure, and that M, N & X are measurement variables of the problem, these assumptions are not strictly necessary; it is sufficient that M & N are cut-variables, and that $\{\Lambda_x\}_{x \in \mathbb{X}}$ partitions $\Lambda(\mathcal{C})$. And even these conditions can be relaxed, as we see in Example 3.

Example 3. *An alternative drug becomes available, resulting in a revised sCEG as in Figure 10.*

Let $W_a = \{w_0\}, W_b = \{w_1, w_2\}, W_c = \{w_3, w_4\}$ and $W_d = \{w_5, w_6, w_7, w_8\}$. Now, unlike W_b, the sets W_c & W_d are not position-cuts as they do not partition $\Lambda(\mathcal{C})$. However, we can still define

$$X(W_c) = \sup_{w \in W_c} X(w),$$
$$X(W_d) = \sup_{w \in W_d} X(w)$$

$X(W_c), X(W_d)$ (although not cut-variables) are both measurable with respect to the sigma field of $\mathcal{C}$, but can, unlike $X(W_a)$ or $X(W_b)$, take zero values, if a patient does not display the symptom.

If we let Λ_1 be the event S *displayed but drug not given*, then we get the sub-SCEG $\mathcal{C}_{\Lambda_1}$ shown in Figure 9, from which we can read the statement

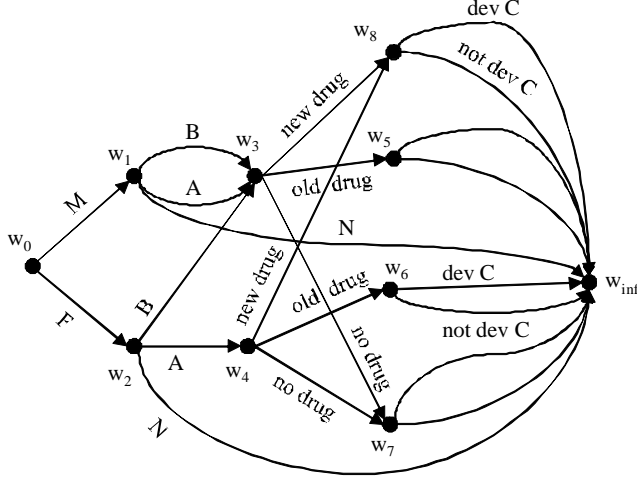$$(X(W_a), X(W_b)) \amalg X(W_d) \mid \Lambda_1.$$

Figure 10: sCEG for Example 3

If we let Λ_2 be the event *old drug given*, then we get the graph $\mathcal{C}_{\Lambda_2}$ shown in Figure 8, and as we have already shown,

$$(X(W_a), X(W_b)) \not\!\amalg X(W_d) \mid \Lambda_2.$$

If we let Λ_3 be the event *new drug given*, then we get a graph $\mathcal{C}_{\Lambda_3}$ which differs from that in Figure 9 only in that the cut-vertex is now w_8, not w_7. We can then read the statement

$$(X(W_a), X(W_b)) \amalg X(W_d) \mid \Lambda_3.$$

Note that $\{\Lambda_i\}_{i=1,2,3}$ here does not partition $\Lambda(\mathcal{C})$.

Clearly we can call $X(W_a)$ & $X(W_b)$ *gender* (X_G) & *symptom* (X_S). If we let $X(w_3)$ & $X(w_4)$ take the values 1, 2 & 3 for the outcomes *no drug, old drug* & *new drug*, then $X(W_c)$ takes the values 0, 1, 2 & 3 for *did not display* S *so did not receive drug, displayed* S *but did not receive drug, received old drug* and *received new drug*. So there is also no ambiguity in calling $X(W_c)$ *drug* (X_D). Taking a similar approach to $\{X(w_i)\}_{i \in \{5,6,7,8\}}$ we find that there is also no ambiguity in calling $X(W_d)$ *condition* (X_C), and (since $X(W_c) = 0 \Rightarrow X(W_d) = 0$) collecting these statements together gives the property

$$X_C \amalg (X_G, X_S) \mid (X_D \neq 2),$$

ie. *condition* is independent of *gender* & *symptom* given that did not receive the old drug.

4 Discussion

Chain Event Graphs were introduced for the representation & analysis of problems for which the use of Bayesian Networks is not ideal. The class of models expressible as a CEG includes as a proper subset the class of models expressible as faithful regular or context-specific BNs on finite variables. Unlike the BN, the CEG embodies the structure of the model state space and any context-specific information in its topology. In this paper we have justified the use of sCEGs for investigating context-specific conditional independence queries of the form $X \amalg Y \mid \Lambda$?, and provided a separation theorem for sCEGs and position variables. The introduction of cut-variables (analogous to BN measurement variables, but more flexible) provides a repertoire of techniques which will enable researchers to tackle a comprehensive collection of conditional independence enquiries on models of asymmetric problems for which the available quantitative dependence information cannot all be embodied in the DAG of a BN.

The research that led to this paper also yieded a number of other questions, some of which are discussed here. The most obvious of these is *Does the* only if *part of Theorem 2 hold if we allow constraints on a CEG's edge-probabilities such as two edge-probabilities being equal?* The short answer is *No*, but the problem is somewhat more subtle than this answer suggests. Some preliminary work on this is described in [40], but a more comprehensive analysis awaits a future paper.

For illustrative convenience the CEGs in the examples in this paper have been constructed in temporal order, but this is not the only valid ordering of a CEG. In [42] for instance, we had a CEG representing a police investigation where the order of events was that in which the police took action or discovered evidence (Extensive Form order [35]). At the simplest level, there are valid reorderings of a CEG in which the cut-variables appear in a different sequence, and there is a set of rules governing when adjacent cut-variables can be *swapped* to produce a different valid ordering. For CEGs depicting models which have a natural product space structure with no context-specific anomolies, these rules are relatively straightforward, but for more general CEGs where we might need to consider swaps of sets of adjacent edges rather than of cut-variables, the rules become very complex. However, it seems fairly certain that if two cut-variables in a coloured CEG are independent then there is a valid reordered pseudo-ancestral CEG of these variables in which the variables are separated by a cut-vertex. We hope to yield more light on this in a future paper.

In [5] we have also looked at infinite CEGs where an individual might come back to (essentially) the same state at some future time point. These problems can be expressed as a CEG analogous to the 2-time-sliced Dynamic BN [19], or as a graph which is no longer acyclic. Both representations involve modification to the rules governing conditional independence structure. This is discussed in [5], but there is an opportunity here for developing CEG semantics further.

Appendix: Proofs

Proof of Theorem 1:

Consider the underlying tree $\mathcal{T}$ of the CEG $\mathcal{C}(\mathcal{T})$. The event Λ corresponds to a union of routes of $\mathcal{T}$. Let $\mathcal{T}_\Lambda$ be the reduced tree consisting only of the vertices, edges & routes that comprise Λ.

If we denote the probabilities of events in $\mathcal{T}_\Lambda$ by $p_\Lambda(..)$, then clearly we require that $p_\Lambda(\lambda) = p(\lambda \mid \Lambda)$.

Once route probabilities in a tree are given, edge-probabilities are uniquely defined. So letting the edge-probabilities in $\mathcal{T}_\Lambda$ be denoted by $\hat{\pi}_e(v' \mid v)$, and letting the route $\lambda \in \Lambda$ be described by its edges as:

$$\lambda = \Lambda(e(v_0, v_1)) \cap \Lambda(e(v_1, v_2)) \cap \cdots \cap \Lambda(e(v_p, v_q)),$$

we have:

$$
\begin{aligned}
p_\Lambda(\lambda) &= p(\lambda \mid \Lambda) \\
&= p(\Lambda(e(v_0, v_1)), \Lambda(e(v_1, v_2)), \ldots \Lambda(e(v_p, v_q)) \mid \Lambda) \\
&= p(\Lambda(e(v_p, v_q)) \mid \Lambda(e(v_0, v_1)), \Lambda(e(v_1, v_2)), \ldots, \Lambda) \\
&\quad \times \cdots \times p(\Lambda(e(v_1, v_2)) \mid \Lambda(e(v_0, v_1)), \Lambda) \\
&\quad \times p(\Lambda(e(v_0, v_1)) \mid \Lambda) \\
&= p_\Lambda(\Lambda(e(v_p, v_q)) \mid \Lambda(e(v_0, v_1)), \Lambda(e(v_1, v_2)), \ldots) \\
&\quad \times \cdots \times p_\Lambda(\Lambda(e(v_0, v_1))) \\
&= p_\Lambda(\Lambda(e(v_p, v_q)) \mid \Lambda(v_p)) \\
&\quad \times \cdots \times p_\Lambda(\Lambda(e(v_1, v_2)) \mid \Lambda(v_1)) \\
&\quad \times p_\Lambda(\Lambda(e(v_0, v_1)))
\end{aligned}
$$

using the Markov property of trees from section 2.2

$$
\begin{aligned}
&= \prod_{e(v,v') \in \lambda} p_\Lambda(\Lambda(e(v, v')) \mid \Lambda(v)) \\
&= \prod_{e(v,v') \in \lambda} \hat{\pi}_e(v' \mid v)
\end{aligned}
$$

If we now let $\mathcal{C}_\Lambda$ inherit the edge-probabilities from $\mathcal{T}_\Lambda$, we have:

$$
\begin{aligned}
p_\Lambda(\lambda) &= \prod_{e(w,w') \in \lambda} \hat{\pi}_e(w' \mid w) \\
&= \prod_{e(w,w') \in \lambda} p_\Lambda(\Lambda(e(w, w')) \mid \Lambda(w))
\end{aligned}
$$

where, without ambiguity, we let $p_\Lambda(..)$ denote the probability of an event in $\mathcal{C}_\Lambda$.

Then

$$\hat{\pi}_e(w' \mid w) = p_\Lambda(\Lambda(e(w, w')) \mid \Lambda(w))$$
$$= p(\Lambda(e(w, w')) \mid \Lambda(w), \Lambda)$$
$$= \frac{p(\Lambda(e(w, w')), \Lambda(w), \Lambda)}{p(\Lambda(w), \Lambda)}$$
$$= \frac{p(\Lambda \mid \Lambda(e(w, w')), \Lambda(w)) \; p(\Lambda(e(w, w')), \Lambda(w))}{p(\Lambda \mid \Lambda(w)) \; p(\Lambda(w))}$$
$$= \frac{p(\Lambda \mid \Lambda(e(w, w'))) \; p(\Lambda(e(w, w')), \Lambda(w))}{p(\Lambda \mid \Lambda(w)) \; p(\Lambda(w))}$$

since $\Lambda(e(w, w')) \subset \Lambda(w)$

$$= \frac{p(\Lambda \mid \Lambda(e(w, w')))}{p(\Lambda \mid \Lambda(w))} \; \pi_e(w' \mid w)$$

Under this edge-probability assignment, no edges in $\mathcal{C}_\Lambda$ are given a zero probability, since each $e(w, w') \in \lambda \in \Lambda$. And no position in $\mathcal{C}_\Lambda$ needs to be split (uncoalesced) in order for us to make this edge-probability assignment.

By construction two vertices in a tree on the same route cannot be in the same position. So consider two vertices in $\mathcal{C}_\Lambda$ which do not lie on the same route. Then the collections of routes (elements of Λ) passing through each of these vertices are disjoint. So we can assign the probability distribution over these routes (in $\mathcal{C}$) in such a way that the conditional joint probability distributions on the subpaths emanating from these two vertices in $\mathcal{C}_\Lambda$ are different. Hence our assignment does not require us to coalesce distinct positions in $\mathcal{C}_\Lambda$.

So position-structure is preserved.

Hence $\mathcal{C}_\Lambda$ is an sCEG, and the set of sCEGs is closed under conditioning on an intrinsic event.

Proof of Lemma 1:

Variables X, Y partition the set of atoms of $\mathcal{C}$, and since $\Lambda \subset \Lambda(C)$, X, Y also partition the set of atoms of $\mathcal{C}_\Lambda$.
Consider arbitrary events Λ_x, Λ_y from $\{\Lambda_x\}_{x \in \mathbb{X}}$, $\{\Lambda_y\}_{y \in \mathbb{Y}}$, and the event $\Lambda_x \cap \Lambda_y$.
Then $p(\Lambda_x \mid \Lambda) = p_\Lambda(\Lambda_x)$ etc., and the statement

$$p(\Lambda_x, \Lambda_y \mid \Lambda) = p(\Lambda_x \mid \Lambda) \; p(\Lambda_y \mid \Lambda)$$

is true if and only if the statement

$$p_\Lambda(\Lambda_x, \Lambda_y) = p_\Lambda(\Lambda_x) \; p_\Lambda(\Lambda_y)$$

is true. If either of these relationships holds for all $\Lambda_x \in \{\Lambda_x\}_{x \in \mathbb{X}}$, $\Lambda_y \in \{\Lambda_y\}_{y \in \mathbb{Y}}$, then so does the other for all $\Lambda_x \in \{\Lambda_x\}_{x \in \mathbb{X}}$, $\Lambda_y \in \{\Lambda_y\}_{y \in \mathbb{Y}}$.
Hence $X \amalg Y \mid \Lambda$ if and only if $X \amalg Y$ in $\mathcal{C}_\Lambda$. $\qquad \square$

Proof of Lemma 2:

1. Consider a single route λ consisting of a subpath $\mu_0(w_0, w)$ between w_0 and w, the edge $e(w, w')$ labelled x ($\neq 0$), and a subpath $\mu_1(w', w_\infty)$ connecting w' to w_∞. Now this route consists of a set of edges and by construction the probability $p(\lambda)$ of the route is equal to the product of the probabilities labelling each of these edges. Moreover, the probability of any subpath of λ is equal to the product of the probabilities labelling each of its edges. So $p(\lambda)$ can be written as the product of the probabilities of three subpaths: $\mu_0(w_0, w), e(w, w')$ and $\mu_1(w', w_\infty)$. Thus:

$$p(\lambda) = \pi_{\mu_0}(w \mid w_0) \, \pi_e(w' \mid w) \, \pi_{\mu_1}(w_\infty \mid w').$$

But the fact that λ utilises the subpath $\mu_0(w_0, w)$ between w_0 and w allows us to completely specify the value of the vector $\boldsymbol{X}_{U(w)}$. By a slight abuse of notation we can represent this as $\boldsymbol{X}_{U(w)} = \mu_0(w_0, w)$.

Consider now the event $(\boldsymbol{X}_{U(w)} = \mu_0(w_0, w), I(w) = 1, X(w) = x)$, which is the union of all $w_0 \rightarrow w_\infty$ routes which utilise the subpath $\mu_0(w_0, w)$ and the edge $e(w, w')$. Then since this is an intrinsic event we can write:

$$
\begin{aligned}
p(\boldsymbol{X}_{U(w)} &= \mu_0(w_0, w), \, I(w) = 1, X(w) = x) \\
&= p(\Lambda(\mu_0(w_0, w)), \Lambda(w), \Lambda(e(w, w'))) \\
&= \pi_{\mu_0}(w \mid w_0) \, \pi_e(w' \mid w) \sum_{\mu_1 \in M_1} \pi_{\mu_1}(w_\infty \mid w'),
\end{aligned}
$$

where M_1 is the set of all subpaths from w' to w_∞. But $\sum_{\mu_1 \in M_1} \pi_{\mu_1}(w_\infty \mid w') = 1$ since all paths through w' terminate in w_∞.

Similarly, for the event $(\boldsymbol{X}_{U(w)} = \mu_0(w_0, w), I(w) = 1)$ we have:

$$p(\boldsymbol{X}_{U(w)} = \mu_0(w_0, w), I(w) = 1) = p(\Lambda(\mu_0(w_0, w)), \Lambda(w)) = \pi_{\mu_0}(w \mid w_0).$$

So

$$
\begin{aligned}
p(X(w) = x \mid \boldsymbol{X}_{U(w)} = \mu_0(w_0, w), I(w) = 1) &= \frac{\pi_{\mu_0}(w \mid w_0) \, \pi_e(w' \mid w)}{\pi_{\mu_0}(w \mid w_0)} \\
&= \pi_e(w' \mid w) = p(\Lambda(e(w, w')) \mid \Lambda(w)) \\
&= p(X(w) = x \mid I(w) = 1).
\end{aligned}
$$

Hence

$$X(w) \amalg \boldsymbol{X}_{U(w)} \mid (I(w) = 1) \qquad\qquad (1)$$

2. If $I(w) = 1$ then $X(w') = 0$ for all $w' \in D^c(w) \cap U^c(w)$, so we can completely specify the value of the vector $\boldsymbol{X}_{D^c(w) \cap U^c(w)}$, and expression (1) implies:

$$X(w) \amalg \boldsymbol{X}_{U(w)} \mid (\boldsymbol{X}_{D^c(w) \cap U^c(w)}, I(w) = 1) \qquad\qquad (2)$$

Moreover if $I(w) = 1$, no further information about $\boldsymbol{X}_{D^c(w) \cap U^c(w)}$ will assist us in predicting the value of $X(w)$. Hence

$$X(w) \amalg \boldsymbol{X}_{D^c(w) \cap U^c(w)} \mid (I(w) = 1) \qquad\qquad (3)$$

Using a result from [11], the expressions (2) and (3) yield the result:

$$X(w) \amalg (\boldsymbol{X}_{U(w)}, \boldsymbol{X}_{D^c(w) \cap U^c(w)}) \mid (I(w) = 1)$$
$$\Rightarrow X(w) \amalg \boldsymbol{X}_{D^c(w)} \mid (I(w) = 1)$$

3. If $I(w) = 0$ then $X(w) = 0$, and no further information about $\boldsymbol{X}_{D^c(w)}$ will assist us in predicting the value of $X(w)$. Hence also

$$X(w) \amalg \boldsymbol{X}_{D^c(w)} \mid (I(w) = 0) \qquad \square$$

Proof of Theorem 2:

1. Sufficient conditions for independence. The sufficient conditions for independence are an almost immediate consequence of similar results for Markov processes, but we include a proof here for completeness.

Consider an sCEG $\mathcal{C}$, and two positions $w_1, w_2 \in V(C) \backslash \{w_\infty\}$ such that $w_1 \prec w \prec w_2$ for some cut-vertex w. By construction $I(w_1) \not\equiv 0, I(w_2) \not\equiv 0$, $I(w) \equiv 1$.

Consider the event $(X(w_1) = x_1, X(w_2) = x_2) \equiv (X(w_1) = x_1, I(w) = 1, X(w_2) = x_2)$ for $x_1 \neq 0, x_2 \neq 0$. This is the union of all routes passing through w_1, utilising an edge $e(w_1, w_1')$ labelled x_1, passing though w, passing through w_2, and utilising an edge $e(w_2, w_2')$ labelled x_2. By analogy with the proof of Lemma 2 we can, since this is an intrinsic event, write:

$$p(X(w_1) = x_1, X(w_2) = x_2) = \sum_{\mu_0 \in M_0} \pi_{\mu_0}(w_1 \mid w_0) \, \pi_e(w_1' \mid w_1) \sum_{\mu_1 \in M_1} \pi_{\mu_1}(w \mid w_1')$$
$$\times \sum_{\mu_2 \in M_2} \pi_{\mu_2}(w_2 \mid w) \, \pi_e(w_2' \mid w_2) \sum_{\mu_3 \in M_3} \pi_{\mu_3}(w_\infty \mid w_2')$$

where M_0 is the set of all subpaths from w_0 to w_1, M_1 is the set of all subpaths from w_1' to w, M_2 is the set of all subpaths from w to w_2, and M_3 is the set of all subpaths from w_2' to w_∞. But $\sum_{\mu_0 \in M_0} \pi_{\mu_0}(w_1 \mid w_0)$ is simply the probability of reaching w_1 from w_0 etc., so this equals

$$= \pi(w_1 \mid w_0) \, \pi_e(w_1' \mid w_1) \, \pi(w \mid w_1') \, \pi(w_2 \mid w) \, \pi_e(w_2' \mid w_2) \, \pi(w_\infty \mid w_2')$$
$$= \pi(w_1 \mid w_0) \, \pi_e(w_1' \mid w_1) \times 1 \times \pi(w_2 \mid w) \, \pi_e(w_2' \mid w_2) \times 1$$

Similarly for the event $X(w_1) = x_1$, we can write:

$$p(X(w_1) = x_1) = \pi(w_1 \mid w_0) \, \pi_e(w_1' \mid w_1) \times 1$$

so

$$p(X(w_2) = x_2 \mid X(w_1) = x_1) = \frac{\pi(w_1 \mid w_0) \, \pi_e(w_1' \mid w_1) \, \pi(w_2 \mid w) \, \pi_e(w_2' \mid w_2)}{\pi(w_1 \mid w_0) \, \pi_e(w_1' \mid w_1)}$$
$$= \pi(w_2 \mid w) \, \pi_e(w_2' \mid w_2)$$

Now consider the event $(X(w_2) = x_2) \equiv (I(w) = 1, X(w_2) = x_2)$. Analogously with above we can write:

$$p(X(w_2) = x_2) = \pi(w \mid w_0)\, \pi(w_2 \mid w)\, \pi_e(w_2' \mid w_2)\, \pi(w_\infty \mid w_2')$$
$$= 1 \times \pi(w_2 \mid w)\, \pi_e(w_2' \mid w_2) \times 1$$
$$= p(X(w_2) = x_2 \mid X(w_1) = x_1)$$

It is straightforward to show that the result also holds for $x_1 = 0$ and $x_2 = 0$.

If w_2 is itself a cut-vertex (with $w_1 \prec w_2$), then we replace $I(w) = 1$ by $I(w_2) = 1$ in the above argument with the same result.

So a sufficient condition for $X(w_1) \amalg X(w_2)$ is that either w_2 is itself a cut-vertex, or there exists a cut-vertex w such that $w_1 \prec w \prec w_2$.

2. Necessary conditions for independence.

Let $X(w_1) \amalg X(w_2)$ (and since $I(w)$ is a function of $X(w)$, $X(w_1) \amalg I(w_2)$ and $I(w_1) \amalg I(w_2)$). Let the set of routes of $\mathcal{C}$ be partitioned into four subsets. Call a route Type A if it passes through w_2, but not through w_1, Type B if it passes through neither w_1 nor w_2, Type C if it passes through both w_1 and w_2, and Type D if it passes through w_1, but not through w_2. Our proof proceeds as follows:

(a) We show that we must have $w_1 \prec w_2$ (ie. the set of Type C routes is non-empty).

(b) We show that every route intersects with every other route at some point downstream of w_0 and upstream of w_∞. If two $w_0 \to w_\infty$ routes share no vertices except w_0 and w_∞, we call them *internally disjoint*. So there cannot be two internally disjoint $w_0 \to w_\infty$ routes in $\mathcal{C}$

(c) We show that there must therefore be a cut-vertex between w_0 and w_∞.

(d) We show that either w_1 is a cut-vertex or w_2 is a cut-vertex, or there exists a cut-vertex w such that $w_1 \prec w \prec w_2$.

(e) Finally we show that if w_1 is a cut-vertex then there must also either be a cut-vertex at w_2 or a cut-vertex w such that $w_1 \prec w \prec w_2$.

(a) Suppose that $w_1 \not\prec w_2$ (and recall that $w_2 \not\prec w_1$). Then $p(I(w_2) = 1 \mid I(w_1) = 1) \equiv 0$. $I(w_1) \amalg I(w_2) \Rightarrow p(I(w_2) = 1) \equiv 0$ $\Rightarrow I(w_2) \equiv 0$. This is impossible by construction. Therefore $w_1 \prec w_2$.

(b) We first show that each Type C route intersects with every other route at w_1 or at w_2 or at some point between these positions.

Let λ_1 be a Type C route, and $\mu_1(w_1, w_2)$ the subpath coincident with λ_1 between w_1 and w_2. If the set of Type B routes is non-empty then let λ_2 be a Type B route which does not intersect with μ_1 (ie. λ_2 and μ_1 have no positions or edges in common).

Consider a distribution P which assigns (1) a probability of $1 - \epsilon$ to every edge of the subpath $\mu_1(w_1, w_2)$, and (2) a probability of $1 - \delta$ to each edge of the route λ_2. Let the number of edges in $\mu_1(w_1, w_2)$ be $n(\mu_1)$ and the number of edges in λ_2 be $n(\lambda_2)$ (where both $n(\mu_1)$ and $n(\lambda_2)$ are finite). Then let

$(1 - \epsilon)^{n(\mu_1)} > 0.9$ and $(1 - \delta)^{n(\lambda_2)} > 0.8$. If λ_2 does not intersect with μ_1 then this is always possible.

Under P, assignment (1) gives us that

$$p(I(w_2) = 1 \mid I(w_1) = 1) \geq (1 - \epsilon)^{n(\mu_1)} > 0.9$$

and $I(w_1) \amalg I(w_2)$ implies that under this P

$$p(I(w_2) = 1 \mid I(w_1) = 0) > 0.9 \;\Rightarrow\; p(I(w_2) = 0 \mid I(w_1) = 0) < 0.1$$

But assignment (2) gives us that

$$p(I(w_2) = 0) \geq p(I(w_1) = 0, I(w_2) = 0) \geq p(\lambda_2) = (1 - \delta)^{n(\lambda_2)} > 0.8$$
$$\Rightarrow p(I(w_2) = 0 \mid I(w_1) = 0) > 0.8 \qquad \text{※}$$

The assumption $I(w_1) \amalg I(w_2)$ is incompatible with the assignments of (1) and (2). But these assignments are always possible if λ_2 does not intersect with μ_1. Hence λ_2 must intersect with μ_1.

Hence each Type C route intersects with every Type B route at some point downstream of w_1 and upstream of w_2. Also each Type C route intersects with every Type A route (at w_2), with every Type D route (at w_1) and with every other Type C route (at both w_1 and w_2).

We now consider routes that are not of Type C. If the set of non-Type C routes is non-empty let λ_3, λ_4 be members of this set which do not intersect except at w_0 and w_∞. Let $\mu(w_1, w_2)$ be a subpath between w_1 and w_2.

From above both λ_3 and λ_4 must intersect with μ. Let λ_3 intersect with μ only at the positions $w_{31}, \ldots w_{3m}$, where $w_{31} \prec \cdots \prec w_{3m}$; and let λ_4 intersect with μ only at the positions $w_{41}, \ldots w_{4n}$, where $w_{41} \prec \cdots \prec w_{4n}$. Without loss of generality let $w_1 \preceq w_{31} \prec w_{41} \preceq w_2$, so that λ_3 could be a route of Type B or Type D, and λ_4 could be a route of Type A or Type B.

Suppose firstly that $w_{4n} \prec w_{3m}$. Consider the subpath $\mu_5(w_1, w_2)$ which coincides with μ from w_1 to w_{31} (if $w_{31} \neq w_1$), coincides with λ_3 from w_{31} to w_{3m}, and coincides with μ from w_{3m} to w_2. This subpath μ_5 does not intersect with the route λ_4. This is impossible since every route in $\mathcal{C}$ intersects with every $\mu(w_1, w_2)$ subpath.

Suppose therefore that $w_{3m} \prec w_{4n}$. Consider the subpath $\mu_6(w_1, w_\infty)$ which coincides with μ from w_1 to w_{31} (if $w_{31} \neq w_1$) and coincides with λ_3 from w_{31} to w_∞; and the subpath $\mu_7(w_0, w_2)$ which coincides with λ_4 from w_0 to w_{4n} and coincides with μ from w_{4n} to w_2 (if $w_{4n} \neq w_2$). Consider also a distribution P which assigns (1) a probability of $1 - \epsilon$ to every edge of μ_6, and (2) a probability of $1 - \delta$ to every edge of μ_7. Let the number of edges in $\mu_6(w_1, w_\infty)$ be $n(\mu_6)$ and the number of edges in $\mu_7(w_0, w_2)$ be $n(\mu_7)$ (where both $n(\mu_6)$ and $n(\mu_7)$ are finite). Then let $(1 - \epsilon)^{n(\mu_6)} > 0.9$ and $(1 - \delta)^{n(\mu_7)} > 0.8$. If λ_3 and λ_4 do not intersect then this is always possible.

Under P, assignment (1) gives us that

$$p(I(w_2) = 0 \mid I(w_1) = 1) \geq (1 - \epsilon)^{n(\mu_6)} > 0.9$$

and $I(w_1) \amalg I(w_2)$ implies that under this P

$$p(I(w_2) = 0 \mid I(w_1) = 0) > 0.9 \;\Rightarrow\; p(I(w_2) = 1 \mid I(w_1) = 0) < 0.1$$

But assignment (2) gives us that

$$p(I(w_2) = 1 \mid I(w_1) = 0) > 0.8 \qquad \text{※}$$

The assumption $I(w_1) \amalg I(w_2)$ is incompatible with the assignments of (1) and (2). But these assignments are always possible if λ_3 and λ_4 do not intersect. Hence λ_3 and λ_4 must intersect.

Hence each Type B route intersects with every Type A, Type B or Type D route, and each Type A route intersects with every Type D route. Also, each Type A route intersects with every other Type A route (at w_2), and each Type D route intersects with every other Type D route (at w_1). So each route in $\mathcal{C}$ intersects with every other route downstream of w_0 and upstream of w_∞.

Hence there cannot be two internally disjoint directed routes from w_0 to w_∞.

(c) To show that this implies the existence of a cut-vertex between w_0 and w_∞, we briefly consider a CEG as a Flow Network where every edge and every vertex (except w_0 and w_∞) has a (flow) capacity of one. Then the maximum flow through the CEG from w_0 to w_∞ must equal the maximum number of internally disjoint $w_0 \to w_\infty$ routes. We can now use Ford & Fulkersons' Max Flow Min Cut Theorem [13]. This theorem applies to networks where only the edges are given capacities, so we replace each vertex $w \in V(\mathcal{C}) \backslash \{w_0, w_\infty\}$ by a pair of vertices w^-, w^+ connected by an edge $e(w^-, w^+)$ with a capacity of one – the only edge emanating from w^- being $e(w^-, w^+)$ and the only edge entering w^+ being $e(w^-, w^+)$.

The theorem states that for a Flow Network with a single source and a single sink, the maximum flow from source to sink equals the capacity of the minimum cut, where cuts pass through the **edges** of the graph (ie. a cut partitions $V(\mathcal{C})$ into two collections of vertices with w_0 in one collection and w_∞ in the other), and the capacity of the minimum cut is the sum of the capacities of the edges which are cut.

So if in our CEG we have no pairwise internally disjoint $w_0 \to w_\infty$ routes, then the maximum flow through the CEG from w_0 to w_∞ must equal one, and the capacity of the minimum cut of the CEG must also equal one. Hence all $w_0 \to w_\infty$ routes must pass through a single edge.

Now this edge may be of the form $e(w^-, w^+)$, in which case w is a cut-vertex; or the edge may be of the form $e(w_a, w_b)$ for $w_a \neq w_b$, in which case both w_a and w_b are cut-vertices. Hence there is a cut-vertex w such that $w_0 \prec w \prec w_\infty$.

This result can also be arrived at by using a corollary of Whitney's [44] Theorem 7 (a result for undirected graphs, sometimes described as the *2nd variation of Menger's Theorem* [25]).

(d) Suppose there exists a cut-vertex upstream of w_1. Then relabel this cut-vertex as w_0 and repeat the argument of (b)(c) to show that there exists a cut-vertex between this new w_0 and w_∞. Since the number of positions in $\mathcal{C}$ is

finite, repeated use of this argument shows us that either w_1 is a cut-vertex or there exists a cut-vertex downstream of w_1. A complementary argument shows that there exists a cut-vertex at w_2 or upstream of w_2.

(e) Suppose w_1 is a cut-vertex, but w_2 is not. Then either (i) w_2 lies exactly one edge downstream of w_1 on every $w_1 \to w_2$ subpath, or (ii) there exists a position w_1^1 ($\neq w_2$) exactly one edge downstream of w_1 lying on a $w_1 \to w_2$ subpath.

(i) We know that $X(w_1) \neq 0$ (since w_1 is a cut-vertex), so if $X(w_1)$ takes a value corresponding to an edge from w_1 to w_2, then $I(w_2) = 1$ and $X(w_2) > 0$; otherwise $I(w_2) = X(w_2) = 0$. So $X(w_2) \,\slashed{\amalg}\, X(w_1)$.　　　※

(ii) If $X(w_1)$ takes a value corresponding to an edge from w_1 to w_1^1, then $I(w_1^1) = 1$; otherwise $I(w_1^1) = 0$. Hence $I(w_1^1)$ is a function of $X(w_1)$. So $X(w_1) \amalg X(w_2) \Rightarrow X(w_1) \amalg I(w_2) \Rightarrow I(w_1^1) \amalg I(w_2)$, and using the argument of (b), (c), (d) above there must be a cut-vertex at w_1^1 or between w_1^1 and w_2.

Therefore there exists a cut-vertex at w_2 or a cut-vertex w such that $w_1 \prec w \prec w_2$.

$\square$

Proof of Corollary 1:

Let $X(w_1) \amalg X(w_2)$ hold for some $w_1 \in W_a$, $w_2 \in W_b$. Then by Theorem 2 either (i) w_2 is a cut-vertex (in which case W_b consists of the one position w_2), or (ii) there exists a cut-vertex w such that $w_1 \prec w \prec w_2$.

Since W_a and W_b are position cuts, this implies that either (i) $w_a \prec w_2$ $\forall w_a \in W_a$, or (ii) $w_a \prec w \prec w_b$ $\forall w_a \in W_a, w_b \in W_b$, and hence (i) $X(w_a) \amalg X(w_2)$ $\forall w_a \in W_a$, or (ii) $X(w_a) \amalg X(w_b)$ $\forall w_a \in W_a, w_b \in W_b$.

Note that $X(w_a)$, $X(w_b)$ pairwise independent for all w_a, w_b does not in general imply groupwise independence, but it does here:
Any event characterised by the expression $\boldsymbol{X}_{W_a} = \boldsymbol{x}_a$ has the form:

$$X(w_a') = x_a \ (\neq 0) \text{ for some } w_a' \in W_a, \quad X(w_a) = 0 \quad \forall w_a \in W_a \backslash \{w_a'\}$$

So

$$p(\boldsymbol{X}_{W_a} = \boldsymbol{x}_a, \boldsymbol{X}_{W_b} = \boldsymbol{x}_b)$$
$$= p(X(w_a') = x_a, X(w_b') = x_b, X(w) = 0 \quad \forall w \in W_a \cup W_b \backslash \{w_a', w_b'\})$$

for some $w_a' \in W_a, w_b' \in W_b$

$$= p(X(w_a') = x_a, X(w_b') = x_b)$$

since $X(w_a') \neq 0 \Rightarrow X(w_a) = 0 \quad \forall w_a \in W_a \backslash \{w_a'\}$ etc

$$= p(X(w_a') = x_a)\, p(X(w_b') = x_b)$$

since $X(w'_a) \amalg X(w'_b)$

$$
\begin{aligned}
&= p(X(w'_a) = x_a, X(w_a) = 0 \quad \forall w_a \in W_a \backslash \{w'_a\}) \\
&\quad \times p(X(w'_b) = x_b, X(w_b) = 0 \quad \forall w_b \in W_b \backslash \{w'_b\}) \\
&= p(\boldsymbol{X}_{W_a} = \boldsymbol{x}_a) \, p(\boldsymbol{X}_{W_b} = \boldsymbol{x}_b)
\end{aligned}
$$

So $\boldsymbol{X}_{W_a} \amalg \boldsymbol{X}_{W_b}$. But $X(W_a) = \sup_{w \in W_a} X(w_a)$ is a function of $\boldsymbol{X}_{W_a}$, and $X(W_b)$ is a function of $\boldsymbol{X}_{W_b}$. Hence $X(W_a) \amalg X(W_b)$.

$\square$

Proof of Corollary 2:

Since Λ is intrinsic to $\mathcal{C}$, $\mathcal{C}_\Lambda$ is a subgraph of $\mathcal{C}$ with $V(\mathcal{C}_\Lambda) \subset V(\mathcal{C})$.

Let W_a in $\mathcal{C}_\Lambda$ be the subset of $V(\mathcal{C}_\Lambda)$ which consists of elements of W_a in $\mathcal{C}$. Then W_a is well-defined on $\mathcal{C}_\Lambda$, as is $X(w_a)$ for any $w_a \in W_a$.

$X(W_a)$ is measurable with respect to the sigma-field of $\mathcal{C}$, so it partitions the set of atoms of $\mathcal{C}$. Since $\Lambda \subset \Lambda(\mathcal{C})$, it also partitions the set of atoms of $\mathcal{C}_\Lambda$, and is well-defined on $\mathcal{C}_\Lambda$ as:

$$
X(W_a) = \sup_{\substack{w_a \in W_a \\ w_a \in V(\mathcal{C}_\Lambda)}} X(w_a).
$$

Hence $p_\Lambda(X(W_a) = x_a) = p(X(W_a) = x_a \mid \Lambda)$, and all necessary terms are defined on $\mathcal{C}_\Lambda$ consistently with their definitions on $\mathcal{C}$.

In $\mathcal{C}_\Lambda$ there exists a cut-vertex w such that $W_a \prec w \prec W_b$, so by Theorem 2, $X(w_a) \amalg X(w_b)$ holds in $\mathcal{C}_\Lambda$ for any $w_a \in W_a \cap V(\mathcal{C}_\Lambda), w_b \in W_b \cap V(\mathcal{C}_\Lambda)$.

Hence by Lemma 1, $X(W_a) \amalg X(W_b) \mid \Lambda$ holds in $\mathcal{C}$.

$\square$

References

[1] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37:3099–3132, 2009.

[2] S. A. Andersson, D. Madigan, and M. D. Perlman. Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28:33–85, 2001.

[3] L. M. Barclay, J. L. Hutton, and J. Q. Smith. Refining a Bayesian Network using a Chain Event Graph. *International Journal of Approximate Reasoning*, 54:1300–1309, 2013.

[4] L. M. Barclay, J. L. Hutton, and J. Q. Smith. Chain Event Graphs for Informed Missingness. *Bayesian Analysis*, 9:53–76, 2014.

[5] L. M. Barclay, J. Q. Smith, P. A. Thwaites, and A. E. Nicholson. The Dynamic Chain Event Graph. Research Report 14-04, CRiSM, 2014. Submitted to *Electronic Journal of Statistics*.

[6] R. R. Bouckaert and M. Studeny. Chain graphs: Semantics and expressiveness. In C. Froidevaux and J. Kohlas, editors, *Symbolic and Quantative approaches to Reasoning and Uncertainty*, number 946 in Lecture Notes in Artificial Intelligence, pages 67–76. Springer-Verlag, 1995.

[7] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian Networks. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

[8] M. Bozga and O. Maler. On the Representation of Probabilities over Structured Domains. In *Computer Aided Verification*, volume 1633 of *Lecture Notes in Computer Science*, pages 261–273. Springer, 1999.

[9] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, 1999.

[10] R. G. Cowell and J. Q. Smith. Causal Discovery through MAP selection of stratified Chain Event Graphs. *Electronic Journal of Statistics*, 8:965–997, 2014.

[11] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41:1–31, 1979.

[12] A. P. Dawid and M. Studeny. Conditional products: an alternative approach to conditional independence. In *Proceedings of the 7th Workshop on Artificial Intelligence and Statistics*, pages 32–40, 1999.

[13] L. R. Ford and D. R. Fulkerson. *Flows in Networks*. Princeton University Press, 1962.

[14] G. Freeman and J. Q. Smith. Bayesian MAP model selection of Chain Event Graphs. *Journal of Multivariate Analysis*, 102:1152–1165, 2011.

[15] G. Freeman and J. Q. Smith. Dynamic staged trees for discrete multivariate Time Series: Forecasting, model selection and causal analysis. *Bayesian Analysis*, 6:279–305, 2011.

[16] S. Hojsgaard and S. L. Lauritzen. Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society, Series B*, 70:1005–1027, 2008.

[17] M. Jaeger, J. D. Nielsen, and T. Silander. Learning Probabilistic Decision Graphs. In *Proceedings of the 2nd European Workshop on Probabilistic Graphical Models*, pages 113–120, Leiden, 2004.

[18] O. Kallenberg. *Foundations of Modern Probability*. Springer, 1997.

[19] K. B. Korb and A. E. Nicholson. *Bayesian Artificial Intelligence*. Chapman & Hall/CRC Press, 2004.

[20] S. L. Lauritzen. *Graphical Models*. Oxford, 1996.

[21] S. L. Lauritzen. Causal inference from graphical models. In O. E. Barndorff-Nielsen et al., editors, *Complex Stochastic Systems*. Chapman and Hall, 2001.

[22] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H. G. Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–505, 1990.

[23] D. McAllester, M. Collins, and F. Periera. Case factor diagrams for structured probabilistic modeling. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 382–391, 2004.

[24] C. Meek. Strong completeness and faithfulness in Bayesian Networks. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 411–441, 1995.

[25] K. Menger. Zur allgemeinen Kurventheorie. *Fundamenta Mathematicae*, 10:95–115, 1927.

[26] D. M. Q. Mond, J. Q. Smith, and D. Van Straten. Stochastic factorisations, sandwiched simplices and the topology of the space of explanations. *Proceedings of the Royal Society of London*, A 459:2821–2845, 2003.

[27] S. M. Olmstead. *On Representing and Solving Decision Problems*. PhD thesis, Stanford University, 1983.

[28] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge, 2000.

[29] D. Poole and N. L. Zhang. Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence Research*, 18:263–313, 2003.

[30] T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.

[31] A. Salmeron, A. Cano, and S. Moral. Importance sampling in Bayesian Networks using probability trees. *Computational Statistics and Data Analysis*, 34:387–413, 2000.

[32] G. Shafer. *The Art of Causal Conjecture*. MIT Press, 1996.

[33] T. Silander and T-Y. Leong. A Dynamic Programming Algorithm for Learning Chain Event Graphs. In *Discovery Science*, volume 8140 of *Lecture Notes in Computer Science*, pages 201–216. Springer, 2013.

[34] J. Q. Smith and P. E. Anderson. Conditional independence and Chain Event Graphs. *Artificial Intelligence*, 172:42–68, 2008.

[35] J. Q. Smith and P. A. Thwaites. Decision trees. In E. L. Melnick and B. S. Everitt, editors, *Encyclopedia of Quantitative Risk Analysis and Assessment*, volume 2, pages 462–470. Wiley, 2008.

[36] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Springer-Verlag, 1993.

[37] P. A. Thwaites. Causal identifiability via Chain Event Graphs. *Artificial Intelligence*, 195:291–315, 2013.

[38] P. A. Thwaites and J. Q. Smith. Evaluating causal effects using Chain Event Graphs. In *Proceedings of the 3rd European Workshop on Probabilistic Graphical Models (PGM)*, pages 291–300, Prague, 2006.

[39] P. A. Thwaites and J. Q. Smith. Non-symmetric models, Chain Event Graphs and Propagation. In *Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 2339–2347, Paris, 2006.

[40] P. A. Thwaites and J. Q. Smith. Separation theorems for Chain Event Graphs. Research Report 11-09, CRiSM, 2011.

[41] P. A. Thwaites, J. Q. Smith, and R. G. Cowell. Propagation using Chain Event Graphs. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 546–553, Helsinki, 2008.

[42] P. A. Thwaites, J. Q. Smith, and E. M. Riccomagno. Causal analysis with Chain Event Graphs. *Artificial Intelligence*, 174:889–909, 2010.

[43] T. Verma and J. Pearl. Causal networks: semantics and expressiveness. In *Proceedings of the 4th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 1988.

[44] H. Whitney. Congruent Graphs and the Connectivity of Graphs. *American Journal of Mathematics*, 54(1):150–168, 1932.